Simulation of initial medical problem-solving : studies on a new measure for the assessment of medical problem-solving ability

SIMULATION OF INITIAL MEDICAL PROBLEM-SOLVING

Studies On A New Measure
For The Assessment Of
Medical Problem-solving Ability

PROEFSCHRIFT:

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof.dr. F.I.M. Bonke,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen op donderdag,
25 mei 1989 om 16.00 uur

door

Erik de Graaff
geboren in 1951 te Amsterdam

promotor: Prof.dr. M.J. Drop
co-promotor: Dr. H.J.M. van Berkel

beoordelingscommissie: Prof.dr. J.J.C.B. Bremer (voorzitter)
Prof.dr. C.P.A. van Boven
Dr. J.G.M. Gerritsma
Dr. H.F. Kraan
Prof.dr. G.R. Norman

& Uitgeverij Thesis, Haarlem
Omslagontwerp: Roelof van Rosmalen
Druk: Krips Repro
ISBN 90-5170-006-7

voor: *Annemaaike, Bregje, Nadine, Marjorie en Thea*

"No problem is so big or so complicated, that it can't be run away from" Charly Brown.

Table of contents page

Something extra is needed to turn research into a thesis. The push to go for that extra I owe very much to my promoters. Riet Drop, professor of medical sociology, project leader of "Follow-up Basisartsen" became a guiding force for the research on measurement of medical problem-solving. Though, as a member of the faculty board, she had little time to spare (which sometimes made it difficult to arrange meetings), she always was able to provide me with extensive criticism on the drafts I submitted. I am very much indebted for those conscientious comments.
Henk van Berkel, member of the department of educational research and development, became involved with the project in 1987. Especially during the last important year of actually writing the thesis his support was most valuable. Riet and Henk, I thank you both for the support in writing this thesis.

Looking back it seems a long time ago that the research presented in this thesis was initiated. It began in 1978 with a proposal for a project for the development of a measure for the assessment of medical problem-solving in addition to the Maastricht Progress Test, written by Maarten Verwijnen and Tjaart Imbos. In 1979 Jan Galesloot (physician) and the present author (psychologist) were assigned to the project.
The association with Jan I remember as one of the most fruitful periods of my life. The basic conception of a new measure for assessment of medical problem-solving emerged from endless, but never tiring discussions between the two of us. We developed a measure with open-ended questions, based on medical case histories, called "Summatieve Evaluatie van Medisch Probleemoplossen" (Summative Evaluation of Medical Problem-solving). Very soon we were undertaking empirical experiments to try out SEMP and even published a book presenting our measure.
Unfortunately the co-operation with Jan ended too early. Despite the enthusiasm for our research accomplishments Jan needed the challenge of practical medical work. After almost two years he left the project, returning to health care practice. To me it seemed the project lost a sense of direction, without the participation of Jan as a person, and as a

medical expert. After working for about half a year at the ungrateful task of finishing the project single-handed, I was offered the opportunity to work as an educational psychologist in the faculty of Health Sciences. Though I did not intend to abandon the medical problem-solving project completely the work on the construction of a new health sciences curriculum quickly occupied almost all of my time. Hence, much research remained uncompleted.

A renewed interest in the measurement of medical problem-solving grew out of my association with the project Follow-up Basisartsen (graduating physicians). This project aimed at evaluating the curriculum of the medical school of Maastricht by means of investigating the achievement of graduates in medical practice. For this purpose reliable and valid measures of medical competence were necessary. In this context SEMP was considered a potentially useful measure. However, more empirical evidence was needed to corroborate the quality of the instrument. With the assistance of two Maastricht graduates, Pieter van de Heijden and Beppie Hellemons-Boode an experiment was conducted, comparing SIMP with other measures of medical problem-solving. Gerry Post, researcher on the Follow-up project, did a wonderful job analyzing the data of this experiment.

When asked to replace Gerry Post and complete the Follow-up investigations, in 1985, I also got the opportunity to continue research with SEMP. The promising results with respect to SEMP were recognized by Geoffrey Norman from McMasters University in Canada, who stimulated the writing of an article in the English language. He also suggested to transform the name of the instrument into "Simulation of Initial Medical Problem-solving" (SIMP). Since this description better fits the instrument, this name will be used from here on. Many thanks, Geoff, for the suggestions and the motivating comments on earlier drafts of the article.

Only after publication of this article in 1987 did I decide to continue the research on SIMP with the purpose of producing a thesis. In a sense it was hard to return to the original track. During that period, my enthusiasm for SIMP was revived in numerous discussions with Yvonne van Leeuwen who joined the Follow-up project for a year, while pursuing a

research course and Thom Lenz who entered the project as a student research assistant.

In the almost ten years that are covered by my research on medical problem-solving many more people have contributed large, or small, but irreplaceable elements. I am grateful to all who have helped in any capacity to realize these studies: the physicians who provided the raw cases, those who helped transforming them into test material, the nurses and physicians who did the scoring, the Maastricht graduates and students that worked on the project and my colleagues of the department of educational development and research who had to listen tirelessly to reports of my progress.
The assistance of mrs. Diane Riksen with the data analysis, was very helpful. Further, I am indebted to the numerous members of the secretary staff who have typed out countless pages of SIMP reports. A special thanks is reserved for mrs. Petry Thiemann, who prepared the final lay-out of this book.
I am grateful for the thorough criticism on the manuscript I received from the referees. Especially the chairman of the committee Prof.dr. J.J.C.B. Bremer was most helpful in pointing out detailed flaws in the text.

Finally, the writing of a thesis is very much a solitary job. My friend Thea was able to support me from a distance. To her and to our children I dedicate this thesis.

1 INTRODUCTION

1.1 Problem-based Medical Education

Sworn by the oath of Hippocrates, physicians are the guardians of health. All patients consulting a physician are entitled to a professionally adequate response their health problems. In order to train physicians who are able to provide such responses, medical education should offer students opportunities to acquire the necessary professional knowledge and skills.

Medical education traditionally is information-intensive. A classical medical curriculum roughly follows a pattern of first teaching basic sciences, each separately, subsequently different medical disciplines, and finally practical training in internships. Students are required to memorize a multitude of isolated facts, before they get a chance to apply their knowledge in practice. Many students will have forgotten a lot of basic knowledge by the time they enter their internships. Moreover, they have learned to rely on authorities to decide for them what is relevant. As a consequence they may have a hard time dealing on their own with situations in which they are confronted with a lack of knowledge and skills.

In the early seventies a new approach to medical education was developed at McMaster University in Canada. In this approach integration and application of knowledge were considered to be of more importance than storing facts by rote learning. Students should start learning from practical medical problems right from the beginning of the study. Therefore, this educational approach is referred to as problem-based learning.

(Spaulding, 1969; Neufeld and Barrows, 1974; Fraenkel, 1978; Barrows and Tamblyn, 1980). Learning is centered on problems from medical practice. By analyzing problems, students acquaint themselves with the problem-solving process of a physician. Also, they have early opportunities to learn to integrate knowledge from different disciplines, related to the same medical problem. By formulating questions with respect to the information they lack to solve a problem, students select their own learning-goals. Trained as independent learners, students from a problem-based curriculum may be expected to be able to identify and fill gaps in their knowledge, also after graduation.

The ability to keep their knowledge and skills up-to-date, is a crucial factor in determining the competence of our physicians to be. The necessity to adapt medical education to changing needs of society is emphasized in a report by the Association of American Medical Colleges' Project Panel on the General Professional Education of the Physician (GPEP) (Muller, 1984). As a result of their analysis of the factors influencing contemporary medical education, the panel arrived at the conclusion that the strong emphasis on factual knowledge in traditional curricula is undesirable. Ongoing advances in medical science have expanded the body of medical knowledge far beyond the comprehension of one single person. Hence, it has become impossible to add all new relevant information to the existing curricula. Medical curricula are not only overcrowded already, but also information is being made obsolete by rapid advances in bio-medical knowledge and technology. Medical practice has branched off into numerous specialties, each requiring years of additional professional training. Since most graduates enlist for further education, the function of the basic medical curriculum has changed, from direct entrance in preparing for medical practice, to providing a general background for further education.

Furthermore, medical practice is affected by demographic changes in society, such as the increase of the elderly population, and the number of people with chronic diseases. Also insights with respect to medical interventions are changing. For instance, there is growing recognition for the influence of environmental factors and life-style on health. Altogether,

it is very difficult to indicate which information learned now, will be useful in practice ten or twenty years from now.
Nevertheless, the GPEP-panel stresses the importance of an uniform basic medical education: "all physicians, regardless of specialty, require a common foundation of knowledge, skills, values and attitudes". With respect to desired adaptations of medical education, the panel formulates a number of recommendations. In a review, on the effectiveness of problem-based medical education compared with traditional curricula, Schmidt and colleagues stipulate that several of those recommendations (like less emphasis on factual knowledge, reduction of scheduled time, and promotion of independent learning and problem-solving), are fulfilled by problem-based medical education (Schmidt et al, 1987). Although the available data were as yet insufficient to draw unequivocal conclusions, the authors suggest problem-based curricula seem to be successful in providing a learning environment, congruent with the general goals of university education. For instance, students in a problem-based curriculum were found to display a more positive motivation towards their study than students in a traditional medical curriculum (De Graaff et al, 1982). Although there is some evidence suggesting that students in a problem-based curriculum master slightly less factual medical knowledge, the difference with students from traditionally curricula disappeared at the end of the study (Bender, et al, 1984; Verwijnen et al, 1988). Further, problem-based learning was demonstrated to be more efficient. Compared with traditional Dutch medical schools a higher percentage of students of the problem-based curriculum in Maastricht graduates in less time (Post et al, 1986).

1.2 The Maastricht curriculum

The founding of a new medical school in Maastricht, in 1972, constituted an excellent opportunity for innovation of medical education in the Netherlands (Basisfilosofie, 1972; Greep, 1979). The medical school of the Rijksuniversiteit Limburg was among the first to adapt the principles of problem-based learning (Schmidt and Bouhuijs, 1980; Schmidt, 1983).

Following the example of the medical faculty of McMaster University, Maastricht puts its students to work together on medical problems presented to them by the staff in small tutorial groups in which they formulate their own learning goals. In addition to the group-work, skills training and practice orientation are scheduled. A large proportion of time is reserved for independent self-study. In response to the epidemiological and demographic shifts in the population (Walton, 1985), community-oriented health care is emphasized the Maastricht curriculum.

The four-year preclinical phase of the curriculum is organized in thematic six-week blocks. Small multidisciplinary teams of faculty members are responsible for the content of the "block books", introducing the theme to the students by means of study tasks. These study tasks are descriptions of patient cases or other phenomena related to the current theme. Working on these tasks the students first analyze the problem and try to solve it by applying their apriori knowledge on the subject. In this process they find out where they lack important knowledge and subsequently formulate learning goals. The acquired insights are discussed in the next session. During these blocks students are trained in the skills laboratory, practicing medical skills which are closely related to the patient problems in their "block books".

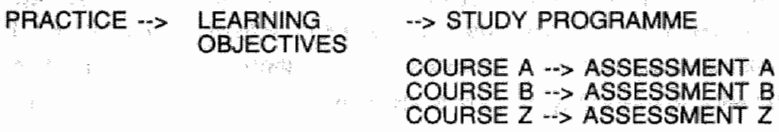
A two-year clinical phase completes the curriculum. In various clinical settings, including a two-month psycho-medical clerkship and three months practice in family medicine, students acquire clinical experience. Due to the freedom of students in selecting their own learning goals, the control of the medical school over the content of students' learning is limited to the offering of study materials as a guideline of study activities and to indirect control by means of assessment procedures. Nevertheless, the emphasis on health care seems to be apprehended by the students. Post and Drop (1988) conducted a study in which the perception of curriculum content among Maastricht graduates and graduates from traditional curricula were compared. They found that Maastricht graduates reported that more attention had been given to subjects like: "primary health care, mental health care, behavioral science and ethical issues". Students from traditional curricula indicated more attention had

been given to biomedical science and clinical medicine. In many cases they even declared that too much attention had been given to these subjects.
The perceived emphasis on primary health care, however, does not seem to have a notable effect on the career choices of graduates. In a study on career preference and career realization of Maastricht graduates, the number of graduates opting for a career as family physician about equaled the number preferring a career as clinical specialist (De Graaff et al, 1987-b). The recent surplus of graduated physicians in the Netherlands, however, may interfere with career-choice and thus obscure the effect of curriculum content.

1.3 Assessment in a problem-based curriculum

Graduation from medical school means one is allowed to practice medicine. In the Netherlands medical schools are authorized to construct their own examinations (within certain limits specified by law). There is no equivalent to the American National Board examination. This places the responsibility on the medical schools for setting their own standards against which medical students are to be judged.
Usually there is no direct connection between the demands of practice and assessment procedures. At a very global level study objectives are derived from medical practice. These study objectives are translated into a study programme outline (a list of disciplines, with or without order of relevance and preferred sequence). Student assessment is related to the contents of this programme. (see figure 1).

figure 1:



As the study programme is divided into separate courses assessment follows this division. At the end of each course students have to take a test. Those who pass can continue with the study, those who fail must either take the course a second time, hand in extra assignments, take a second chance at the test, or drop out. The Dutch methodologist and educational psychologist De Groot (1973) characterizes this type of examination system as a "steeple chase". The barriers that are put in the way of smooth study progress cause unnecessary study delays (for instance, if it is not allowed to follow a course before satisfactory completing the preceding one a student has to wait - sometimes up to a year - to be able to do so, when he or she has failed that test).

At the start of the Maastricht curriculum student assessment was designed following the lines described above. The six-week thematic blocks were regarded as courses, and at the end of each block a test was taken. Since problem-based learning encourages students to formulate their own learning objectives within the block themes, the content of what they learn is not as clearly defined as in a regular course. As any test covers only a small sample of the actual learning objectives, it is not surprising that students try to forecast which parts will be tested, and concentrate their study activities accordingly. This results in test-directed learning, instead of problem-based learning.

To counter this unwarranted effect of block-tests a different approach to assessment appeared to be necessary (Wijnen and Van der Vleuten, 1985). At the medical faculty of the Rijksuniversiteit Limburg the results of block-tests were discarded for assessment purposes. (They were still taken, but only as a source of information to the students; later on the block-test were again taken into consideration, as supporting evidence). For student assessment a special instrument was developed, the so called Maastricht Progress Test (Verwijnen et al, 1982). This is a large test at graduate level, sampled from the total medical body of knowledge. Test-directed learning is avoided, because there is no direct relationship between the test and the recently completed course (see figure 2).

figure 2:

PRACTICE --> LEARNING OBJECTIVE --> STUDY PROGRAMME
--> ASSESSMENT AT END LEVEL

All students of every level have to take a Progress Test four times a year. Each test consists of 200-300 true/false items, with a question mark option. Naturally, students in their first year lack a lot of knowledge necessary to answer the questions, and will have to use the question mark option more often than students of later years. For assessment purposes the scores of students are related to the mean-score of their cohort.

Research has indicated that the Progress Test is a fairly reliable measure. Furthermore, it has been demonstrated that successive scores through the years take the shape of growth curves, indicating the growth of knowledge towards the end-level of the curriculum (Verwijnen et al, 1982).

Though, satisfactory in many respects, the Progress Test is focussed on the assessment of factual knowledge. And since physicians need more than just factual medical knowledge to function in health care practice, the test does not completely cover the learning objectives of the curriculum. Practical medical skills, for instance, cannot be assessed by means of the Progress Test. Hence, a separate skills test has been developed (Van der Vleuten and Van Luyk, 1985; Bouhuijs et al, 1987).

Another domain that is not adequately covered by the Progress Test is usually labelled "medical problem-solving". In addition to the Progress Test another measure was therefore necessary. In 1977-1978 an attempt was made to introduce Patient Management Problems (PMP) for this purpose. PMP's are paper-and-pencil simulations of patient-physician interactions, inspired by early work of Rimoldi (1961). Several variations of this method were developed, for instance: "Diagnostic Management Problem (DMP), Sequential Management Problem (SMP) (Galofrè, 1974; McGuire, 1974; Martin, 1975). Basically all these instruments follow the same pattern: a case description is followed by series of optional ans-

wers, allowing the candidate to select a path to solve the problem. In subsequent steps, the whole process of medical problem-solving is encompassed.

The results of the try-out in Maastricht were unsatisfactory (unpublished). The construction of two PMP's proved to be difficult and time consuming. Next the production of the test material with invisible text, to be made visible with special latent image developers caused many problems. Further the results appeared to depend heavily on the method of scoring. Also, first year students were handicapped by their lack of understanding of medical terminology in the case-histories.

Support for the decision to discard PMP's as a measure of medical problem-solving can be found in the literature. After initial enthusiasm for this new measurement technique more and more studies appeared, casting doubt on the validity and reliability. Repeatedly low correlations were reported, both among PMP's (McGuire, 1976; Elstein et al, 1978; Page and Fielding, 1980; Newble et al, 1982), and with concurrent measures (Mura et al, 1976; Lamont and Hennen, 1979). These low correlations raise serious doubt with respect to the validity of decisions based on such lists.

Further, serious problems with the scoring method were exposed. The effect of cueing by the optional answers was shown to increase the scores markedly (Goran, 1973; Martin, 1975; Newble et al, 1979; Norman and Feightner, 1981). Also, it proved to be very difficult to assign score weights, and to find a balance among different answering strategies (McGuire, 1974; Schwabbauer, 1975). Finally the PMP is expensive, both in actual costs (the printing procedure with latent images is expensive), and in production time.

Therefore, another measure needed to be found which could be used in addition to the Progress Test. This thesis deals with the search for such a measure.

1.4 Plan of the book

This thesis is largely based on a series of articles describing the construction of SIMP and subsequent research. As yet, two of these articles have been published, in Medical Education and in Medical Teacher respectively. A third article has been accepted for publication in Medical Education. The text of the articles has been adjusted in order to prevent the redundancy which is inherent in a series of articles on the same subject. The indulgence of the reader is asked when redundancy to some extent still occurs.

As a first step in the development of SIMP, the concept medical problem-solving was analyzed. The results of this analysis and the methodological background of test development are presented in Chapter 2.

Next, the operationalization of medical problem-solving by means of SIMP is described in Chapter 3, based on the article "Simulation of Initial Medical Problem-solving: a test for the assessment of medical problem-solving" (De Graaff, 1988).

Since the instrument consists of case descriptions followed by an open-ended question, the answers must be marked by raters. Both with respect to the objectivity of measurement, as from a practical point of view, this might be seen as a disadvantage. As objectivity is an important condition for measurement, it receives special attention in Chapter 4. Two studies were devoted to the investigation of the agreement among raters. A study on the reliability of scoring by six nurses is described in Chapter 4, paragraph 2. A study, comparing the ratings by nurses, with ratings by experienced physicians, is presented in paragraph 3 of chapter 4.

Objectivity of measurement is a condition for the usefulness of an instrument. There are, however, more factors determining the instrumental utility of a measure. Chapter 5, based on the article: "Validation of a new measure of clinical problem-solving" (De Graaff et al., 1987a), describes an experiment, investigating the reliability and validity of SIMP, within the framework of Generalizability theory. Also, scores on a SIMP test are compared with several concurrent measures.

In Chapter 6 the results of the research with SIMP so far and the usefulness of SIMP as a measure of medical problem-solving are discussed.

2 MEASUREMENT OF MEDICAL PROBLEM-SOLVING: methodology of instrument development

2.1 Introduction

Assessment of learning results is probably as old as learning itself. Originally, tutors judged pupils according to their own standards. Individual or small scale tutoring allowed close observation and questioning. Sometimes apprentices had to demonstrate proficiency by performing a special assignment (like the medieval master test).

Educational measurement as we know it today, however, did not evolve before the end of the nineteenth century. At that time scientific interest focussed on individual differences. When the insight grew that humans do not only differ on physical attributes but also with respect to personality traits or ability to learn, attempts were made to measure those psychological traits. Measurement of educational results followed in the trail of research on intelligence.

Scientific measurement differs from the earlier assessment procedures in several aspects. A broad definition of measurement is: "the assignment of numerals to objects or events, according to rules" (Kerlinger, 1973, p. 426). According to the dictionary, to measure is: "to ascertain the extent, degree, quantity, dimensions, or capacity of, by a standard" (Webster's New International Dictionary, second edition, 1957). These definitions imply that measurement is a quantitative process of expressing results in terms of numbers. In order to do so, there has to be a standard, an "as constant as possible" unit of measurement to give meaning to the numerical expression (Noll, 1965, p. 7). Educational measurements, however, are relative. There are no absolute units of measurement for

possession of knowledge or skills that are comparable to measures from the physical domain, like length, weight, or temperature (Noll, 1965, p. 10).

The main goal of educational measurement is to reflect learning results. According to the classical approach to educational measurement, both curriculum content and test content must be derived from educational goals (Bloom, et al. 1971; Ebel and Frisbie, 1986). Usually, curricula are composed of a number of courses. At the end of each course an achievement test is administered to the students.

Problem-based learning, however, presents an extra challenge to the designers of assessment methods. Because students have considerable freedom in choosing their own learning goals, there is sizable variation in what they are actually learning. Also, problem-based learning is integrative, which entails that elements from different scientific disciplines have to be studied at the same time. Traditional end-of-course tests cannot cope with these problems. Assessment methods that allow for a high degree of variability in learning outcomes (both quantitatively and qualitatively) do not exist (Wijnen, 1984).

However, assessment methods in a problem-based curriculum should meet these extra demands. At the same time they must also fulfil the requirements demanded of any educational measurement. As Neufeld (1984) points out, the same properties that are commonly described in the measurement literature should apply for assessment tools in problem-based programs.

Standards for educational measurement are derived from psychological test development. An outline of the methodological principles that are applied in the development of a new measure for the assessment of medical problem-solving is presented in the following chapter. As main reference, the standard work on methodology by the psychologist De Groot (1969) is utilized.

2.2 Defining medical problem-solving

The first step in the development of a new measure is analysis of the concept-to-be-measured. In a preliminary analysis of the concept "medical problem-solving", the international literature was studied (Galesloot and De Graaff, 1980). In the search for a definition of medical problem-solving repeatedly phrases were encountered like: "Clinical problem-solving is the ultimate goal of medical education" (Palva, 1974), or "Physician's clinical competence depends in great part on their abilities to identify and solve patient problems" (Vu et al, 1984). Such phrases denote the importance generally attached to medical problem-solving. A clear definition of medical problem-solving, allowing the distinction of problem-solving from other aspects of clinical competence, was, however, not found.

Actually, the state of affairs could be characterized as a terminological chaos. Terms like "medical problem-solving", "clinical problem-solving", "clinical reasoning", "medical thinking", are sometimes used as synonyms. In different publications, however, the same terms appear to convey quite different connotations. For instance, Boshuizen and Claessen (1982) pointed out that medical problems can be regarded as patient's problems as well as physician's problems. Such a discrepancy has consequences for the approach to medical problem-solving, especially in deciding when a problem is solved satisfactory.

Evidently medical problem-solving is hard to define. These definition problems could be explained by the fact, that medical problem-solving is a hypothetical concept. The distinction between empirical concepts and hypothetical concepts is a gradual one and is related to the level of abstraction (De Groot, 1969, p. 65). Empirical concepts, like "length" or "weight", are directly related to observed phenomena. Their function is to summarize observations, which they completely cover. Hypothetical concepts are often designated as constructs. They assume the existence of a phenomenon, object or process, that cannot be observed directly, but may serve to explain observed phenomena (McCorquodale and Meehl, 1948, p. 110). Concepts like "black hole" and "anti matter", are examples

of hypothetical constructs from physics. As most processes inside the human mind cannot be observed directly, psychology is exceptionally rich in hypothetical constructs. Psychoanalytical concepts for instance, like "super-ego" and "libido" originate from psychoanalytical theory and cannot be verified by observations. Personality traits like "dominance" and "test anxiety", may be derived from observed phenomena, but cannot be observed directly.

Hypothetical constructs usually have a broader connotation, than is evident from observations. This is referred to as the "surplus-meaning" of hypothetical constructs (Reichenbach, 1938). Especially, since the surplus-meaning is often generated by the metaphorical character of hypothetical constructs, confusion among different definitions of the same concept may result.

The construct medical problem-solving is composed of the terms "medical", "problem" and "solving". Problems are usually associated with difficulties or puzzles. Solving the problem means removing the difficulty or finding the correct answer or solution. Especially the latter is associated with reasoning and thinking. The same associations apply for medical problem-solving. Both, the removing of medical difficulties (patient complaints) and the process of reasoning involved in finding a solution for difficult cases, may rightfully be called medical problem-solving.

However, both approaches are inappropriate for assessment purposes. In the first case the solving of a medical problem would depend primarily on the outcome of the problem-solving process, that is the recovering, improving, or satisfaction of the patient. This approach denies the possibility, that it is not caused by incompetence of the physician when the patient keeps having complaints. The second approach emphasizes the ability to find ingenious solutions for complicated cases. As this is by definition out of the ordinary, this kind of problem-solving does not happen very often in regular medical practice. And future physicians should be tested against situations they may typically expect, not against exceptional incidents.

As a result of the analysis of the concept medical problem-solving the following working definition was formulated (De Graaff and Galesloot, 1981):

> *"Medical problem-solving is the total of activities a physician displays, in reacting on a situation in medical practice".*

This definition concentrates on performance in practice, emphasizing observable medical behavior instead of internal mental processes of the physician. Possession of relevant knowledge is of course necessary, as well as the ability to combine data and to draw conclusions from them. The medical profession, however, is strongly oriented towards practice. In many urgent cases a physician has to act first and can only stop to think about the how and why afterwards. Therefore, correctness of medical actions is regarded as an indication of the ability to solve medical problems.

The emphasis on actions appears to be in line with views presented at the first Cambridge Conference on Directions in Clinical Assessment: "'Clinical actions' were defined as activities undertaken by a clinician in the process of diagnosing and managing a patient's problem(s). We believe the definition should include all aspects of gathering data, making diagnostic and therapeutic decisions and providing treatment and follow-ups - which include patient education. The definition excludes the clinical reasoning process underlying overt clinical actions" (Wakeford et al, 1984, p. 29).

In a recent article Norman (1988) elucidates this same point by emphasizing the distinction between problem-solving and solving problems. In this view "problem-solving" is related to higher order cognitive skills, mental strategies or a reasoning process. From the results of two decades of research Norman concludes that expertise is not characterized by the possession of such general skills. The ability to find adequate responses to patient problems of all sorts, including those with a routine character is referred to as "solving problems". In that case the ability to solve medical problems depends on the actions a physician undertakes

in the process of diagnosing and managing patient problems: gathering additional data, making diagnostic and therapeutic decisions, providing treatment and planning follow-up. The choice of correct actions at each of these steps is largely determined by medical knowledge and practice experience which is in accordance with the schematic representation of clinical competence, presented by Neufeld and Norman in their book on Assessment of Clinical Competence (Neufeld and Norman, 1985). They regard medical problem-solving as one of the elements contributing to clinical competence and they define it as the ability to apply knowledge. This view appears to support the decision to discard the reasoning process in the working definition of medical problem-solving. The labelling "problem-solving" is, however, widely associated with the connotation of a mental process. Following Norman (1988) it would now seem more appropriate to identify the intended construct as "the ability to solve medical problems".

2.3 Operationalization

The working definition of medical problem-solving, presented in the preceding paragraph, is very broad. All medical behavior is included, without differentiation into types of action. Further specification is necessary in order to be able to classify individuals according to their degree of problem-solving ability. In other words, to transform the construct into a variable.

An objective instruction for the activities or operations, that allow such a transformation is called an operational definition (De Groot, 1969 p. 85). With most empirical concepts, it is comparatively easy to formulate operational definitions. Defining length as the distance between two points, amounts to the same as comparing the distance between two points with a standard of length. The operationalization covers the concept-as-intended completely.

In many cases, however, it is impractical or even impossible to utilize an operationalization that completely covers the concept-as-intended. Any concept may be operationally defined in several ways, each partially

covering the original "concept-as-intended". Take, for instance, the numerous tests constructed to measure intelligence. De Groot (1969, p. 67) argues that the freedom to choose an operationalization of a concept or construct is inherent in the freedom of conceptualization in science. Especially with hypothetical constructs associated with ambiguous empirical phenomena it may be difficult to specify the non-operationalized surplus-meaning. The choices in the process of operationalization should, however, be made explicit.

An essential aspect of medical problem-solving (in the above presented working definition) is the generation of responses in practice situations. Hence, practice conditions should be represented in an operationalization of this construct. The objective of developing a measure for large scale testing limits the choice to test material that is easy to handle, can be standardized and produced in large numbers. In medical examinations written case histories are regularly used as simulation of practice conditions (for instance with PMP's). The written case history of a PMP is, however, followed by a series of optional choices. From the beginning of the project an open ended question was felt to be more appropriate in representing the generation of responses (De Graaff and Galesloot, 1981).

The advantages of optional choices as an easy and objective method of scoring are evident and explain their popularity. Suggestions that objective tests are more superficial and less realistic than essay tests are refuted by Ebel and Frisbie (1986, p. 163), who maintain that those criticisms do not apply to good objective type tests. Also the suggestion sometimes made by educators, that objective type tests have an impoverishing effect on education could not be substantiated by research findings (Wesdorp, 1979). In a review on the influence of testing on teaching and learning, Frederiksen (1984) cites several studies indicating that it is more difficult to write objective test items for domains like mental skills than for factual knowledge, but that it is not impossible to do so. Moreover, he also mentions several carefully controlled experiments demonstrating an almost perfect correlation between objective type tests

and open-ended parallel versions, which suggests that the test format that is most easy to administer is preferable.

With a test called "Formulating Hypotheses" (Frederiksen and Ward, 1978), however, low correlations were found between the original open-ended format and a later constructed objective parallel form. As an explanation they suggest that the test aimed at evoking responses to ill defined problems. Only the free response format appeared to require the broad search for relevant ideas and information stored in memory, which is necessary to provide such answers.

"Formulating Hypotheses" aims at measuring skills involved at an early stage of problem-solving and does not elicit cognitive activities involved in the remainder of the problem-solving process. Since the generation of ideas, suppositions, or tentative hypotheses, which becomes manifest in the actions that a physician undertakes directly after encountering a patient, seems to be a similar sort of task, the results reported by Frederiksen and Ward can be regarded as support for the choice of open-ended questions in the operationalization of medical problem-solving.

In the interaction between a physician and a patient there are numerous feed-back loops. New information is requested several times, balanced against already available data and possible diagnoses, before a final conclusion is reached. Simulations of this whole process tend to become rather complex, as is manifested by the work of Gerritsma and Small (1988) who constructed four Interactive Patient Simulation (IPS) that contained about 1600 items.

Supported by findings from research into the cognitive aspects of medical problem-solving, which suggest that the first brief period of time after an encounter with a patient might be critical to the ensuing problem-solving process (Elstein et al, 1978; Cutler, 1979), instrument development was focussed on the initial stage of problem-solving.

The situation of a physician encountering a patient was simulated by means of written case-histories, followed by an open-ended question. These further steps of the operationalization of the ability to solve medi-

cal problems with a measurement procedure designated Simulation of Initial Medical Problem-solving (SIMP) are described in chapter 3.

2.4 Objectivity

The instructions for the transformation of a concept into a variable should be objective. The purpose of scientific activity is to gain knowledge of the object of study. If the results of a study are distorted by subjects conducting the study, inadequate, or even fallacious knowledge is gained. In assessing medical problem-solving, the ability of candidates is the object of interest. A measurement procedure should represent just this ability. Ratings of performance in practice, for instance, are exposed to subjective bias of the judges.

According to De Groot, (1969, p. 163): "the general requirement of objectivity, implies that the investigator must act as 'objectively' as possible, that is, in such a way as to preclude interference or even potential interference by his personal opinions, preferences, modes of observation, views, interests, or sentiments". This general direction states "as objective, as possible", because subjective influence cannot completely be eliminated. For one thing, science itself is a human activity. In the social sciences, human processes are the object of study, and humans are the subjects performing it. Linschoten (1964) labels this phenomenon "zelfbetrokkenheid" (self-involvement) of psychology. In many instances, it causes the object of study to be associated with personal or political viewpoints.

Subjective influences may be encountered in all stages of research. The consequences of the requirements of objectivity, however, may vary according to the phase of scientific activity (De Groot, 1969 p. 167). Preferences of the investigator, for instance, usually play a part in the selection of object of study. Also, freedom of conceptualization, as discussed above, may lead to differences in approach. Hence, some subjective influence in this phase of research cannot be excluded, and in fact may be justified. In this phase of research, the requirement of objec-

tivity primarily demands that a researcher is explicit about the choices made.

With respect to the operationalization of a construct, objective instructions should be available for all operations (observation, registration, categorizing, ordering, computation, etc.) necessary to assign a value to the variable in individual cases. Subjective influence can be greatly reduced by the use of machines. However, the risk exists that technical perfection will be achieved at the expense of the relevance of the measurement (De Groot, 1969, p. 176).

When it is not possible to achieve a sufficiently relevant objective operationalization by mechanical procedures, human judges may be employed. In such cases, the system of judgement should be specified as much as possible. If the instructions for the judges are explicit and clear enough to be carried out by "clerks", the measurement may still be designated as objective (Guilford, 1954). The degree of objectivity can and should be verified empirically by comparing the results of different judges. Hence, in judgmental procedures, the criterion of inter-subjectivity or inter-judge reliability takes the place of the objectivity requirement (De Groot, 1969, p. 222).

As a consequence of the choice for the open-ended test format, raters will be part of the procedure for measuring medical problem-solving. In order to assess problem-solving ability, the responses must be compared with a standard, in one way or another. As in most medical situations several equally correct approaches are feasible, the correctness of a response cannot be verified easily by comparing it with a single standard answer. A multiple standard, requiring at least some degree of interpretation, necessitates the employment of judges. Investigation of the inter-rater reliability, to assess the objectivity of the judgmental procedure, thereby, becomes an important objective of this thesis (see chapter 4).

2.5 Instrumental Utility

Once the objectivity of an operational definition is established satisfactory, the value of the variable can be investigated. Since, the term

"value" may cause confusion, De Groot (1969, p. 245.) introduces the concept "instrumental utility". De Groot distinguishes three aspects of instrumental utility that need to be questioned:

1. Is the resulting variable an acceptable representation of the concept-as-intended?
2. Is the measure sufficiently specific?
3. Is the measurement sufficiently efficient?

The first aspect is usually indicated as the validity of a measure, the second is referred to as reliability, and the third as efficiency or usability. In the following paragraphs these three aspects will be elaborated.

2.5.1 Validity

The first question of instrumental utility refers to the validity of the measure. In a broad sense, validity means the degree to which the results of a measure are really the results of interest (Kerlinger, 1973, p. 457). Validity is influenced by many factors and can be investigated in many different ways, resulting in a multitude of types of validity. In a review, Van Berkel (1984) lists no less than 77 different types of validity classified in five broad categories, and a rest category: I. Criterion oriented, II. Content related, III. Referring to the original trait or construct, IV. From experiments, V. On appearances, VI. Rest.

In research on medical problem-solving the credibility or face validity, belonging in the fifth category, has received much attention. In order to be credible, an instrument must evidently appear to measure what it intends to measure. Paper-and-pencil, or computer-based simulation, like Patient Management Problems (PMP), Modified Essay Questions (MEQ), and also SIMP, fulfill this criterion on face value by representing situations from actual practice. In methodological literature, however, the term face validity is often used to indicate an appearance of validity, sometimes even mockingly called "faith validity" (Drenth, 1971, p. 207). Though it is important for the acceptability of a measure that candidates believe it is valid, "experienced realness" is not enough to sustain validity. At least in one respect test situations always depart from reality:

candidates are aware of being tested. Usually, they are able to accept a certain degree of departure from reality as part of the test situation. The validity of the test should, however, be substantiated by analyses of data in relation to the measurement goal. For instance the measurement properties identified by the first three categories, referring to a criterion, the content and the construct respectively.

Prediction of future performance is often seen as the ideal approach to validity. Performance in practice, the ultimate criterion is, however, difficult to establish. The majority of the problems that patients present to their physicians can be solved in many different ways. And with the exception of real mistakes, many different approaches are medically acceptable. In other words, there exists no generally accepted "golden standard" for the solving of medical problems in practice.

Kane (1982) argues that the preference for predictive validity is not reflected in practice, because it is usually very difficult to obtain adequate criterion measures. He therefore suggests that licensure examinations should be regarded as measures of specific abilities that are critical for professional practice. Neufeld (1984) suggests that construct validity should be used. Construct validity, however, is difficult to assess. It is derived from the relation of a variable to other constructs and variables, and its position in a theoretical framework. At the beginning of instrument development a sound theoretical framework was, however, not available. Support for the construct validity can be derived from the analysis of differences between groups of known different competency levels (divergent and convergent validity).

With respect to the "notion of a general problem-solving process" Norman (1988) concludes that the results of research do not support the existence of a general skill to be applied in a variety of situations. Recent cognitive research on medical problem-solving indicates that differences in performance, between experts and novices, might be explained by differences in the structure of knowledge (Claessen and Boshuizen, 1985). In the mental representation a physician constructs of a patient problem "enabling conditions", like sex, age and occupation seem to play an important part. Medical experts were demonstrated to

be more effective in interpreting such contextual information. (Hobus et al, 1987; Hofstra et al, 1988). Experienced physicians do not necessarily possess more knowledge, but they seem to have easier access to their knowledge base, and thereby they are more able to apply their knowledge in complex practical situations.

SIMP does not aim at measuring a clinical reasoning process, but concentrates on medical actions. Easy access to knowledge and experience with similar cases might very well explain the ability to come up with the right actions. A study on the relation between SIMP and a cognitive measure (Claessen and Boshuizen, 1985) which could have enlightened this matter, however, failed, because the cognitive measure was not able to discriminate between students at different levels. The investigations of validity that are reported in this thesis are limited to correlations with concurrent measures of medical competence.

2.5.2 Reliability

The second aspect mentioned of instrumental utility, is the specificity or reliability of measurement. Traditionally, the reliability of a measure is investigated by repeating the measurement. The less the difference between successive measurements, the more reliable the measure. Most behavioral variables, however, are unstable. Therefore, differences between measurements may also be caused by real changes in the measured factor.

In the classical theory of reliability, precision and stability are distinguished as separate aspects of measurement and treated independently. Precision is related to the amount of information contained by a measurement result: "an instrument measures the more precisely, the more relevant information, on the average, one outcome provides" (De Groot, 1969, p. 262). Hence, precision is a function of the degree of differentiation, allowed by the scale on which the measurement takes place. Grades, ranging from A through E, supply more information, than "pass" or "fail" decisions. As a rule, the degree of differentiation of a measurement scale, should match the purpose of measurement.

Stability is estimated by the replication of a measurement. True score is approximated by the mean of a series of replications. When actual replications are not possible, true score is estimated from the internal consistency of the measure (regarding elements of a measure as replications). The influence of systematic error, however, cannot be demonstrated if all aspects are replicated at the same time. Hence, partial replications, varying one relevant aspect at the time, are used to estimate systematic error (for instance: replication of an experiment with different raters, in order to assess rater reliability). This results in a plethora of different reliability coefficients.

Generalizability theory provides a framework allowing treatment of different aspects of reliability within one design (Cronbach, et al, 1972). The question of reliability is rephrased as a question of generalizability: to what extend is it possible to generalize from an observed score to an universe of scores? The variance of observed scores is broken down into different sources of variance, by means of analysis of variance (Mitchell, 1979; Thorndike, 1982; Brennan, 1983). Depending on the universe one wants to generalize to, the variance of facets (items, raters, sessions, etc.), may be defined, either as contributing to true score, or as part of error score. Reliability is expressed as a quotient of variances and is to be treated as an estimate of the correlation between observed scores and the specified universe.

Taken into consideration that all factors simultaneously influencing reliability are confounded in an individual score, a generalizability coefficient represents a more truthful estimate of measurement reliability. Next to the studies focussing on inter-rater reliability (in order to secure the objectivity of the judgmental procedure) the generalizability of the measure of medical problem-solving was investigated (see chapters 4 and 5).

2.5.3 Efficiency

Validity and reliability are in a sense conditions sine qua non. If a measure is not valid, it simply does not tell us what we want to know, and if it is not reliable, it cannot even begin to tell us what we want to know.

The consequences of deficiencies in efficiency may seem to be less severe, affecting costs rather than outcomes. Actually however, the efficiency of an operationalization is directly related to the degree of reliability and validity that can be realized.

The internal efficiency of an instrument has to deal with the question how effectively and economically the measurement is organized with respect to the research goal (De Groot, 1969, p. 284). This amounts to the determination of an empirical variable with an optimal value for validity and reliability, rather than a maximum value. A simple example is test length. Reliability increases with the number of test items. After reaching a certain number of items, however, the gain in reliability decreases fast and the test may become too long to be of practical use.

Both reliability and validity of measures of medical problem-solving have presented cause for concern (Neufeld and Norman, 1985). Therefore, an optimum with respect to efficiency is hard to find. In some cases it would require days of testing to obtain minimal acceptable reliability (Norman, 1987). SIMP is designed to allow the administering of a fair sample of cases within reasonable testing time. As a consequence the credibility or face validity may have suffered to some extend. This balance between reliability and validity will be discussed in chapter 6.

External factors, like the intended population or the persons that should handle the test may be decisive in determining efficiency. A point of concern with respect to the efficiency of SIMP as operationalization of the ability to solve medical problems, is the use of open-ended questions. The scoring of answers by judges is laborious and expensive. As the optimum with respect to efficiency, will shift in the direction of lower costs, when large numbers of students have to pass the test regularly, implementation of SIMP will have to be restricted to small scale applications.

2.6 Conclusion

Medical problem-solving is hard to define. Firstly, several different connotations prevail. Next, several alternative ways of operationalization are

possible. Emphasizing medical actions, rather than the cognitive process of problem-solving, has the advantage of a direct relation for medical practice. The variety of correct medical responses in actual practice, however, obstructs completely objective measurement. Even, when limited to family physician practice situations, the universe of relevant situations is practically infinite.

The objectivity of the operationalization can be improved, by simulating practice situations. Simulations allow better control over the test content and offer the opportunity to confront each candidate with the same situations. Also, it becomes possible to examine and reexamine rating procedures.

In practice different physicians act differently in the same kind of situation. Open-ended questions allow candidates this same freedom in formulating their response. Also the ability to formulate a response is emphasized, rather than the ability to choose one that looks right. However, the scoring of open-ended questions must be guarded carefully against subjective influences of the judges. It may also affect the efficiency of the measure.

Operationalization of a complex construct, like medical problem-solving, may be described as a series of steps, limiting the original concept-as-intended, balancing to find the optimum of objectivity, validity and reliability. The next chapter deals with the first step, presenting an outline of the development of SIMP as a measure for assessment of the ability to solve medical problems.

3 SIMULATION OF INITIAL MEDICAL PROBLEM-SOLVING: a test for the assessment of medical problem-solving[1]

3.1 Introduction

Actual performance of physicians in health-care practice is an essential feature of clinical competence. The ability to gather data from the patient by history and physical examination, to integrate this information into a diagnostic formulation, to select appropriate investigations, and to institute efficacious management is referred to as medical problem-solving (Norman and Feightner, 1981).

Several investigations illustrate that the relationship between what a physician knows (relevant to a case), and what he actually does in practice is less than perfect (Goran et al, 1973; Rethans and van Boven, 1987). Medical problem-solving abilities are hypothesized to be responsible for the difference and a great effort was made to develop reliable and valid assessment techniques for these abilities. An excellent review is presented in the book on Assessment of Clinical Competence edited by Neufeld and Norman (1985).

The conclusions of this book are not enthusiastically phrased. A conclusion by Norman is that "still no consensus exists about precisely what is clinical competence or clinical problem-solving". Looking back into history it is easy to point out mistakes. By now we can easily see that the labeling of the ability of physicians to act in practice situations is at least unlucky. 'Problem-solving' has a surplus connotation of higher mental processes of reasoning and thinking. In psychological research

[1]Adapted from De Graaff, E. (1988) Simulation of Initial Medical Problem-solving: a test for the assessment of medical problem-solving. Medical Teacher, 10, 1, 49-55.

the problem-solving process is investigated in situations where factual knowledge plays no part, like abstract riddles or match puzzles (Newell and Simon, 1972). Medical knowledge, however, is an essential factor with respect to the ability to respond correctly to problems in health care practice. Therefore, assessment of medical problem-solving cannot be separated completely from medical knowledge.

3.2 Operationalization of medical problem-solving

Representation of medical practice constitutes the foundation of measurement of medical problem-solving. Though the relation to practice is important this does not mean that the best instrument necessarily corresponds to actual practice in the closest way. In order to close the gap between performance and the assessment of that performance we need to decide how to value different elements of performance.
As the performance gets more complex, it becomes more difficult to control the assessment procedures. As a consequence direct observation in practice, the real thing itself, suffers from all kinds of judgement bias (Wakefield, 1985).
The test-situation can be standardized by simulating practice conditions as Barrows (1971) did by employing simulated patients. However, the development of reliable judgment procedures remains difficult (Norman, 1985-b), both with respect to construction of test material as in administering real-life simulations are time consuming.
Paper-and-pencil simulations, like the Patient Management Problem (PMP), which were discarded earlier for application along side the Maastricht Progress Test (MPT), are easier to administer, but obviously depart further from reality, as the written case-description is followed by a series of questions, or optional answers. In order to simulate the interaction between physician and patient sophisticated objective scoring techniques have been developed, like the latent image system (McGuire and Babbot, 1967; McGuire et al, 1972; Skakun and Mclaughlin, 1978), or computer simulations where the optional answers are hidden by the program (Schumacher et al, 1975; Feightner and Norman, 1978). The

early enthusiasm for the objective score-technique of the Patient Management Problem (PMP) was, however, countered by a series of studies questioning the validity and reliability of the measure (Norman and Feightner, 1981; Feightner, 1985).

Again most problems are caused by the scoring. The system of assessing weights to different sections was demonstrated to affect the outcomes (Bligh, 1980). Also, forced-choice items were shown to elicit significantly more correct answers than an open-ended format, an effect attributed to cueing of the optional answers (Newble et al, 1982). Although this study was criticized on methodological grounds (Galesloot et al, 1981) it can be regarded as an indication that much effort is required for the construction of high quality optional choice tests. Another criticism comes from an educational point of view. With ill-structured problems like patient complaints, optional answers can not cover all possible correct responses. Solutions that differ from the optional answers are not rewarded. As a consequence, students are not stimulated in the development of their own approach. When they get familiar with the specific bias of an instrument, they may even learn to restrict themselves to the presenting of expected answers.

Apparently these problems could at least partly be prevented by using open-ended questions instead of optional answers. An example of such a measure is the Modified Essay Question (MEQ) where a case history is followed by a series of open-ended questions (Knox, 1975). Feedback on the progress of the case is provided after each answer. A recent publication, however, signals that the same kind of problems like those with PMP's emerge when this instrument is used over a longer period of time (Feletti and Smith, 1986).

The basic difficulty in constructing a scoring-system for written problem-solving tests is caused by the lack of uniformity between different physicians in the solving of patient problems. This is illustrated by a simulation study with 16 general practitioners and 16 internal medicine specialists (Gerritsma and Smal, 1982).

All physicians completed two Interactive Patient Simulations (IPS). This simulation method enables a physician to ask questions to the patient,

indicate physical examinations, etc. The answers or outcomes are provided by the test conductor, with the aid of a card system. The conformity among the physicians, expressed as point bi-serial correlation ranged from .05 (anamnesis) to .20 (physical examination) for the general practitioners, and from .20 (anamnesis) to .44 (laboratory) for the internal medicine specialists. This low conformity indicated that in particular the general practitioners differ markedly in their approach of a patient problem.

Though this finding is neither worrying nor surprising for any experienced physician, it poses a severe problem for test construction. To what standard should students be judged when experts do not agree? Basically there are two ways to handle this problem. One is to stress the importance of correspondence to reality. Since it is impossible to give a complete and detailed prescription of an ideal action sequence, we then have to rely on expert judges, accepting that their personal bias will influence scores. The other way is to simplify the situation emphasizing characteristic elements and accepting a greater distance with reality. The choice between these two approaches depends on specific circumstances and goals. From the assessment perspective, simplifying seems the more appropriate solution. Not only to gain control over the assessment procedures, but also to allow for more variation of case-material within one test. Especially when the test must be administered regularly to large numbers of students, the construction and production of test material should be fairly easy.

The problems with the reliability and validity of PMP's seem to be at least partly caused by the lack of consistency among cases (Norman et al, 1985). An explanation for the low consistency among cases, is that cases should be regarded as single items in a test. Since consistency of a test is partly a function of the number of items, a low consistency simply indicates that a substantial number of cases is necessary for reliable measurement. McGuire cites a study by Templeton and colleagues who found that "an audit of at least 10 records was required to get a relatively stable estimate of an individual residence's performance in a particular setting" (Templeton et al, 1979 on cit. in: McGuire, 1985).

Simulation of the complete process of problem-solving for each case takes much time (over half an hour for one PMP or interactive MEQ at least). Therefore it is virtually impossible to include a substantial number of different cases within one test session. (Spreading over different sessions makes it easier for the students but makes it no less time consuming). The most outstanding, complicating and time consuming factor in problem-solving tests which are based on case-histories, is the element of interaction. To work through a case from beginning to end one needs to be supplied with information depending on the actions taken.

These problem can be averted by concentrating on the first action directed impulses of the physician, instead of trying to simulate the complete encounter between physician and patient. Research on the nature of medical problem-solving supports this idea that the first brief period of time after an encounter with a patient might be critical to the ensuing problem-solving process (Elstein et al, 1978; Norman, 1985-a). Further, experts and novices differ in their ability to recognize essential features in a case-history (Grant and Marsden, 1987). Those physicians who come up with the correct diagnosis among their first hypotheses are usually able to reach the final conclusion by collecting additional information. Those who do not have the correct hypothesis in the beginning seldom reach the correct solution in the end (Barrows et al, 1982). So for purposes of measurement it seems justified to leave out the interaction and concentrate on the initial actions.

3.3 Simulation of Initial Medical Problem-solving (SIMP)

An important aspect of the operationalization is the representation of medical practice. Medical practice, however, is varied. Physicians work in different specialist settings, and meet a heterogeneous population of patients. The ability to solve problems in one setting is not necessarily representative for the problem-solving in another setting. Since it is virtually impossible to cover the whole domain, the test construction was limited to the domain of general practice (family physician). In the

Netherlands, patients first present their complaints to their family physician. If necessary, they are then referred for specialist treatment in hospital.

Consequently, a general practitioner has to deal with all sorts of complaints, without having to go too deep into specialistic details. Both the variety and the ambiguousness of the complaints a general practitioner encounters, are instrumental in eliciting "problem-solving behavior", as defined. Also, general practice is the best approximation of situations a medical student is expected to be able to deal with at the end-level of the medical school.

It is desirable to use an open-ended question format to avoid the effects of cueing and to be able to reward alternative correct solutions. Furthermore, time spent on one case should be limited, so that a sufficient number of cases can be administered within a session. Finally, an instrument designed to assess the ability to solve medical problems need not cover all of the interaction between a physician and a patient, but should focus on the brief period of time after encountering the problem.

The instrument that was designed following these principles appears to be very simple on the surface. "Simulation of Initial Medical Problem-solving (SIMP)", can be characterized as an essay examination based on short case histories which are followed by the (open) question: "what would you do as a physician in this situation?" The instruction emphasizes that the reactions should be short and concise, based on the first impressions of the available data. It is stressed that completion of a case should take no more than about 5 minutes. The descriptions of case-histories are brief (not exceeding 200 words). The information presented covers personal characteristics and medical data. Some contextual clues, like family back-ground or occupation may be included. Students should, however, not be confused with too much irrelevant information. An example of a SIMP test with two cases and appertaining answering models is presented on page 33. More examples of cases have been published (in the dutch language) as means for self-evaluation (Galesloot and De Graaff, 1981).

Example of a SIMP test

Test instruction

Instruction

In the following test, a number of medical situations will be described.

In each situation you should imagine yourself in the position of a physician who is confronted with the case. You will have to point out what type of reaction, according to you, would be appropriate in that particular case.

This reaction could include anything. Depending on the situation and the available data, you would for instance take one or more of the following steps:

- gather additional information,
- confirm or reject certain assumptions,
- undertake certain (therapeutic) steps.

The clue to the problem is, to indicate what in the described situation is a good procedure. For instance, if in one situation important information is lacking, it is important to be aware of it, and to indicate how to obtain the lacking information; or if in another situation a conclusion or diagnosis can be reached, it is important to make clear which (therapeutic) steps you would choose to take next.

Response

Write down your reaction (short and concise) in the space available, under the written situations.

You need not work on one situation for more than 5 minutes.

Try to react to all situations, even if you are not sure.

Basic question for each problem-situation:

WHAT WOULD YOU DO AS A PHYSICIAN IN THIS SITUATION

Good luck!

SIMP-case 1

Situation: Family physician, home call

Man, 24 years old, factory worker, unmarried. Lives with his parents; unstable family-situation; father is unemployed; mother ill; four adult children living at home.

The patient is known to you to be a heavy smoker and drinker. Since a few months he is complaining of recurrent attacks of upper abdominal discomfort with pain and sickness; no clear cause has been identified. An x-ray investigation of the stomach revealed nothing abnormal.
The relatives have told you on the telephone that the patient is seriously ill. While doing Christmas shopping in a busy shop, he began to feel unwell. He complained of sickness and a constricting sensation in the throat.
You arrive at the patient's home; he sits on a chair, and says he feels dizzy, he looks pale and grayish; you do not smell any alcohol. You observe that he has a fast superficial breathing pattern. The radical pulse is normal, 96/min.

Response:

SIMP-case 2

Situation: family physician, consultation room

Woman, 22 years old, not married, lives with her parents, nursing aide in old people's home, youngest of 5 children.
The G.P. sees her every now and then with complaints of hyperhidrosis*.
Complaint now: since three days sick and vomiting, everything comes back. Stools normal, no abdominal pains. She is not feeling very ill, on the other hand she is not feeling able to work.
When asked whether there has been anything special recently, the patient relates that for three weeks she has been in charge of a nursing department of the old people's home, because the person normally in charge went on holidays. "The old people looked down at me as a youngster, they did not accept any instruction from me". The complaints started directly after this period, when the head of the department had returned from her holiday (three days ago).
You don't notice any abnormality in her physical appearance.

* Hyperhidrosis: excessive sweating

Response:

Scoring-model case 1

DATA GATHERING

Anamnesis (S)

x Has he ever had anything similar before?
x Does he feel pain somewhere, if so, what type of pain? Is the pain radiating, continuous or intermittent?
x Did he vomit? What did the vomitus look like?
x Was there any cause for his collapse? Did it start suddenly? Has he been unconscious?
x Did he have any alcoholic drink today? Other intoxications?
x Is he treated with drugs for anything? Any other ailments?
x Did he have normal defecation? Signs of melaena?

Physical examination (O)

x Examination of chest.
x Vital signs.
x Examination of abdomen.

Diagnoses (A)

x Hyperventilation syndrome.
x Intoxication (alcohol, medicine, other drugs).
x Cardiac insufficiency.
x Respiratory insufficiency.
x Upper abdominal disorder: affection of liver or gallbladder, gastric or duodenal ulcer, pancreatitis.

Plan (P)

x Interruption of hyperventilation attack by talking and reassuring, instruction administration of a sedative or by making him breathe in a bag.
x Education of the patient about the hyperventilation syndrome and how to deal with it.
x Providing continuing contacts to discuss underlying problems.
x Ordering x-rays of stomach, duodenum and gallbladder.
x Have the patient in hospital for further observation.
x Order laboratory tests: HB, WBC, liver tests, urine examination.

Scoring-model case 2

Anamnesis

- What is actually bringing her to the G.P. What does she expect him to do?
- What is her own opinion about the complaints. Can she give more information about possible underlying problems. What may have provoked the present complaints?
- Further information about her working environment.
- Has she had similar complaints at other times? Is she unsure of herself in other situations as well?
- Detailed questions about her physical complaints, concerning digestive tract: what is the appearance of the vomitus? how often does she vomit? how is her appetite? how are bowel movements? did she perhaps use wrong food?
- Other questions like: is urination normal? has she felt feverish? did she take any drug treatment already?

Physical examination

- Deliberately omitting a physical exam.
- Carrying out a physical exam, whether global or detailed.

Evaluation

- Psychosomatic disorder. Stress due to work problems. Personal problems with assertiveness. Nervous breakdown.
- Organic cause like gastroenteritis, food poisoning, gastritis.
- Beware of danger of fixation towards somatic complaints.

Plan

- Continuation of further contacts in order to open discussion about personal problems.
- Drug treatment for vomiting and nausea.
- Discussion about her working environment. Are solutions available.
- Indication of psychotherapeutic approach by the G.P. Try to give her insight into her ways of reacting to stress. Advise to attend to training in assertiveness.

The question "What would you do as a physician in this situation?" is truly open. Not only in the sense that the answer has to be written down instead of choosing options, but also in the sense that many different correct answers are possible. The respondents are directed into picking up clues from the case description which warrant further actions. Those actions may entail the seeking of new information by posing questions, physical examination and requesting laboratory tests. Also among the first impressions can be tentative hypotheses or diagnoses and sometimes indications of a treatment plan (this looks like something for psychotherapy; I need to see this patient back in about a week).

The test instruction does not impose any structure. The respondents are free in the formulations of their answer, only limited by the information presented and by the allotted time. When respondents complain that they need to know the result of their first actions in order to decide what to do next, they are instructed to restrict themselves to those actions they feel imminent on the basis of the presented information. They are, however, free to formulate conditional answers (depending on the outcome of the physical examination I would refer this patient for specialist treatment).

To minimalize rater bias with the marking of SIMP, scoring-models are constructed for each case. The scoring-models are based on a discussion of the case by a team of four to five experienced physicians. Descriptions of elements of a correct response are formulated as items in a checklist. In order to facilitate scoring the scoring-models were organized along the pattern of the SOAP-scheme (Weed, 1969). The respondents are not obliged to structure their answers according to this pattern, but it helps the raters in classifying elements of the answer. The actual scoring is done by raters who check whether the items of the scoring-model are represented in the answer of a candidate.

With these scoring-models no value judgement is demanded from the raters. The answer of a candidate is just checked for items that correspond to descriptions in the scoring-model. Therefore raters do not have to be experts in the field themselves. When raters act like clerks

who inform by perform a task the scoring procedure may be regarded as objective (Guilford, 1954).
Research with SIMP has shown that this method of scoring results in high inter-judge reliability (see chapter 4). Support for concurrent validity was derived from correlation with a global judgement of a simulated patient encounter (De Graaff, et al, 1987-a; see also chapter 5).

3.4 Guidelines for construction

Following the procedures outlined below, experienced physicians can construct SIMP-cases quite easily (Van Leeuwen et al, 1987). As a first step, cases must be collected, selected and adjusted.
The raw material can be based on experiences in practice. Especially cases are suitable, that are open in the sense that there is not just one diagnosis which has to be found. A case which depends primarily on one or two critical actions can only distinguish between those who come up with the right answer and those who do not. Such cases are of course relevant, but there is no reason why they cannot be transformed into good objective test items. SIMP aims at the broad category of ill-defined medical problems, that can be approached with many different strategies. Before it can be used in the test the material must be edited. The problem description should at least include:
- practical circumstances of the situation,
- personal information about the patient, social background,
- medical history (as usually available),
- actual complaint or reason for consulting.

Inclusion of medical observations or interpretations should be avoided, since these may provide clues to implied solutions of the problem. Description of the complaint is best phrased 'in words of the patient'.
A crucial role in the construction is reserved for a panel of experts (physicians with a broad experience in the field). They first comment on the case-descriptions. It is crucial that they can recognize the case as characteristic in relation to their practice experience. An example is the reaction one panel-member gave after reading the case of a 55 year old

taxi driver with lower back pain: "I know this man, I know his taste in food and drink, I know to what kind of pub he goes, I know how he spends his afternoon off, I have treated hundreds of patients like him".
As a next step the panel discusses their own responses to the case.
The test constructor compiles the notes of this discussion and transforms them into a scoring-model by the selection of indicators of correct actions. Whenever necessary he looks for formulations that cover similar actions. For instance, any remark or question that refers to complaints of a patient about pain, is scored under the item "specifying questions about pain").
Next the items are categorized according to the SOAP-system (Weed 1969). The grouping of items in these four categories is also used for the assigning of weights. Arbitrarily each item is counted as one point. The panel decides about the categorizing of the items and the relative weight assigned to the categories (for instance 5 points for Data collection, 2 for Evaluation and 3 for Management and Plan).
In a final stage, the panel reviews both case-and scoring-model. Time spent on the construction of one case can be estimated at about eight hours (four hours for a constructor and four times one hour for panel members).

3.5 Conclusion

A clear advantage of SIMP over other written tests for medical problem-solving, like the PMP and the MEQ, is the short time needed to complete a single case. This makes it possible to administer three to six times as many different cases in a test session of the same length. For measurement purposes this is important. The earlier mentioned problem of low correlations among cases cannot be solved by perfecting detailed measurement procedures within a case. Since the subsequent steps in handling a case usually are dependent on each other, someone who has the right idea in the beginning does well on the whole case. So in order to differentiate between those who are competent and those who are not, it is sufficient to concentrate on a few relevant factors.

An important advantage of the relatively short answering time per case, is that a substantial number of different cases can be included within one test-session. As a consequence of the usually low correlations among cases a considerable number of cases is necessary to attain sufficient reliability.

It might seem to be a disadvantage that the correspondence to reality (often referred to as "face validity") is diminished because the interaction between physician and patient is not represented in SIMP. This might give the impression of a conflict between validity and reliability. Whereas close correspondence to reality is an important factor with respect to the acceptability of an instrument its actual relation with validity is less clear. The validity of an instrument is the extend to which it reflects what it intends to measure. What an instrument can possibly reflect is evidently limited by its reliability. Validity is determined by the content of a test and by comparison with criterion measures. As to the first SIMP again has an advantage because a larger number of cases allows for a better representation of the domain, and the construction of SIMP-cases is comparatively easy.

4 OBJECTIVITY

4.1 Objectivity and Inter Rater Reliability

In any measurement procedure, the primary interest is in information about the measured object(s) and not about the subject performing the measurement. A measurement result may be called objective to the extend that it does full justice to the object of measurement. Absence of subjectivity as a disturbing factor is an important characteristic of objectivity (De Groot, 1969, p. 163).

Particularly, the employment of judges as part of the measurement procedure may result in subjective contamination of the measurement outcomes. A sure way to preclude subjective influences in the scoring of tests is by means of machine operated scoring procedures. Often, the term "objective test" is reserved to denote uniquely tests that can be scored mechanically. Actually, the objectivity of a measure does not depend on the question format (Ebel and Frisbie, 1986, p. 100). The crucial characteristic is whether it is possible to identify correct answers objectively. This implies that it must be possible to define the correct answer by a singular statement or figure.

A difficulty with such responses to situations in medical practice, is that there usually exist several equally correct formulations of a response. Therefore, it is not possible to define a correct answer by presenting one example of a correct response. As a consequence, the traditional objective type of tests, requiring candidates to choose a response from a number of optional answers, are limited to testing whether someone is able to recognize the correctness of the presented approach. The ability

to generate an alternative strategy is not rewarded, and may even act against the candidate.

Open-ended questions as operationalization of medical problem-solving, allow for individually different responses. Because of the impossibility to define the correct response singularly, however, the decision whether a given response is correct or not becomes a matter of interpretation. Human judges must decide whether an answer is correct or not, which introduces the risk of subjective rater bias.

The objectivity of the scoring procedure of SIMP was enhanced by the development of scoring-models. Instead of representing an ideal answer pattern, these scoring-models consist of elements of correct approaches to the problem. Thereby the task of the judges is reduced to recognizing the presence of such elements in the response. However, the formulation of an element will seldom match the exact phrasing of a candidate. Hence, some room for interpretation is left.

The question is, to what extent scores are affected by subjective bias. If different raters come to the same conclusion, the person of the rater evidently does not affect the measurement results. The agreement among raters, or the interrater reliability reflects the objectivity of the scoring.

Raters are only one of the sources of measurement error. This error, however, adds up to the measurement error that open-ended questions and objective type tests have in common.

The two studies reported in the remainder of this chapter focus primarily on the reliability of the scoring of SIMP-answers. The first study analyzes the scoring by nurses, in the second study part of the material is re-scored by physicians. An analysis of the reliability of SIMP as a function of the number of cases and raters is described in the last paragraph.

4.2 A test of medical problem-solving scored by nurses

4.2.1 Introduction

Scoring with the SIMP scoring-models does not involve a value judgement. It is simply a matter of marking those items of the checklist that correspond to elements in the written answer.
An indication for the reliability of these scoring-models was found in a study were data of SIMP were analyzed by means of generalizability theory (De Graaff et al, 1987-a; see also chapter 5). The variance attributed to the scoring of four physicians was estimated zero. It is possible, however, that the ratings of these physicians were influenced by their general impression of the performance of the candidates (a so called "halo-effect").
Since medical expertise is not required for the scoring, it was hypothesized, that non-medical experts should also be able to produce reliable scores. If that is the case, agreement among raters will have to be attributed to the scoring-models. Though medical expertise is not required, knowledge of the medical terminology is needed to recognize synonymous expressions or abbreviations. Nurses, who are familiar with medical idiom, both by their para-medical training and by their work experience in medical environment, fulfill this requirement.

4.2.2 Methods

In a try-out study (De Graaff and Galesloot, 1982) two SIMP-tests consisting of 12 and 10 cases respectively were administered to a group of 25 fifth-year medical students, during their general practitioner residency (PMOH). In this three months practical training period students spent half their time in a family physician's practice. Two times a week groups of eight to ten students meet for two-hour sessions at the university, discussing family medicine. The SIMP-tests were taken as part of the formative evaluation (feed-back to the students), half way and at the end of the course. The cases and scoring-models were constructed in

cooperation with the PMOH-tutors, experienced family physicians. A list with the subjects of the 22 cases is presented in table 1.

table 1: Short description of the SIMP-cases

Case	
1.	Baby, eight weeks, regularly vomiting.
2.	Woman, 40 years, anti-conception pills forgotten.
3.	Boy, 9 years, returning from summer camp, bad appetite, swollen belly.
4.	Man, 36 years, known with stomach complaints, now black diarrhoea, loss of weight.
5.	Man, 45 years, fever, recently returned from tropical country.
6.	Man, 22 years, pain in the chest and breathing problems after sporting.
7.	Woman, 45 years, overweight, hypertension, uses anticonception pill.
8.	Boy, 8 years, stammering.
9.	Woman, 82 years, feverish, coughing, social problems.
10.	Man, 43 years, chronic aspirin user, fever, pain in the scrotum.
11.	Man, 24 years, alcoholic, feinted in a shop.
12.	Man, 30 years, amateur football player, sport injury.
13.	Man, 55 years, psycho-social problems, recurring lower back pain.
14.	Woman, 22 years, problems at work, vomiting and morning sickness.
15.	Woman, 22 years, possibility of venereal disease.
16.	Child, 2,5 years, breath problems.
17.	Man 23 years, acute ear pain, fever, loss of hearing.
18.	Man 73 years, trembling, nervousness, possibility of parkinson.
19.	Woman, 20 years, forgotten to take anti-conceptive.
20.	Boy, 6 years, fever, possible meningitis.
21.	Woman, 80 years, cardiac complaints, dementia, epileptic insult.
22.	Woman 58 years, acute pain in upper belly, dark urine.

The answers were scored by two raters (a physician and a psychologist) and reported to the students. Analyses of these results showed on the whole satisfactory agreement between the raters and a moderate correlation among cases (De Graaff and Galesloot, 1982). The raw answer

material of this study was used in an experiment with the purpose to investigate the hypothesis that para-medically trained persons could act reliably as raters. Since the time allotted on both occasions had been too short, several subjects had not completed all cases. For the scoring study, 500 answers were available. In order to remove the effect of differences in handwriting these answers were typed out.

Six nurses with at least 10 years practical experience were hired for the scoring experiment. They were instructed by a member of the project team (a physician) in about half a day. Each nurse received a copy of the answers and a set of scoring-models/checklists. The actual task of scoring the 500 answers took about 25 hours, about three minutes per answer on average, and resulted in 500 completed checklists with 15 to 25 items, a total of about 10,000 items for each rater.

Assessment of the reliability of raters has received much attention in behavioral sciences. Hence quite a number of different methods of analysis are available to the researcher. (Landis and Koch, 1975; Hartmann, 1977; Kratochwill and Wetzel, 1977; House et al, 1981).

Which method is appropriate in a particular case depends on the purpose of the study and the assumptions one is prepared to make about the data.

According to the classical concept of reliability a test score is partly composed of "true score" and partly of uncorrelated "error score".

$$X_{obs} = X_{true} + X_{error}$$

The error term, which might be positive or negative, is the sum of all influences on the score that are defined as irrelevant. So the effects of the specific test questions, the condition of the subjects, raters, etc. are combined in one random error factor.

Reliability within the classical framework is defined as the ratio of true score variance and observed score variance:

$$R_{xx} = \frac{\text{var true}}{\text{var obs}}$$

Since true variance by definition cannot be observed it has to be estimated statistically.

Generalizability theory provides a multi-facet model which allows for the identification of different sources of measurement error (Lindquist, 1953; Cronbach et al, 1972; Mitchell, 1979; Thorndike, 1982). The design may include items, persons, raters, trials sessions, whichever variable is of interest for the study. By means of analysis of variance the contribution to the total variance of each facet is estimated.

Next, reliability can be computed by dividing the variance of interest by the variance of interest plus the error variance.

$$R_{xx} = \frac{\text{var interest}}{\text{var interest} + \text{var error}}$$

Within this model the composition of the error term can be defined to suit the purpose of the study.

In the study at hand there were persons, tests, cases, items (within cases) and raters as sources of variance. The purpose of the study is to determine the reliability of the raters. The case score is taken as main unit of analysis, since scores are reported at this level. Case scores are probably more reliable than item scores, just as session scores are more reliable than trial scores (Hartmann, 1977). The slight freedom of interpretation of the scoring-models could cause some disagreement at the item-level, which is compensated at the case level.

The problem of assessing rater reliability can be viewed as a special case of the one-facet generalizability study (Fleiss and Cuzik, 1979). Shrout and Fleiss (1979) review the use of the Intra-Class Correlation (ICC) for reliability studies in which n-targets are rated by k-raters. They present a formula which gives an estimate of the reliability of the scoring of one rater randomly selected from a population of raters.

$$ICC = \frac{BMS-EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n}$$

ICC = Intra Class Correlation
BMS = Betweengroups Mean Squares
EMS = Error Mean Squares
JMS = Judges Mean Squares

The result of this calculation can be entered into the Spearman-Brown formula to determine the number of raters needed for a specified reliability.

Next, exploratory analysis was carried out to identify sources of unreliability. In a multi-facet generalizability study this would show up in the interaction terms. Since several subjects in our study did not complete all cases, however, this design is not appropriate. Instead correlational analysis was carried out. The product-moment correlation is a straightforward measure of association between two variables. To find out whether there were systematic differences among the raters correlations were computed both at the level of case-scores and at the level of item-scores. From the average correlation among raters per case both cases with relatively low agreement and raters with systematic low agreement could be identified.

4.2.3 Results

First a two way analysis of variance was performed on the scores of the 500 answers by the six nurses, with both targets and raters as random factors. The results of this analysis are presented in table 2.

table 2: Analysis of variance

source	sum of squares	d.f.	mean square	prob	variance component
mean	101640.48	1	101640.48		33.82
answ.	21425.85	499	42.94	.00	6.97
raters	670.61	5	134.12	.00	.27
error	2755.06	2495	1.10		1.10

When entered in the formula for the Intra Class Correlation (see above) these data produce the following coefficient:

$$ICC = \frac{42.94 - 1.10}{42.94 + 5 \,.\, 1.10 + 6(134.12 - 1.10)/500} = .83$$

This value may be interpreted as an estimate of the correlation of the scoring by one rater, drawn at random from a population of nurses, with the mean of scoring by that population. This value is high and indicates a good inter rater-reliability indeed (a value of .80 is usually given as a criterion for reliability).

Since the study employed six raters it is possible to compute the reliability of the combined rating by all six nurses as well. Shrout an Fleiss (1979) present a formula for the Intra Class Correlation adapted to this case:

$$ICC = \frac{BMS - EMS}{BMS + (JMS - EMS)/n} = \frac{42.94 - 1.10}{42.94 + (134.12 - 1.10)/500} = .97$$

The reliability of the combined scores of six nurses is almost perfect. We may therefore conclude that nurses are very well able to score SIMP reliably.

Despite the high overall reliability, it still might be possible that scoring-errors are concentrated in a few cases and/or raters. Correlations between all pairs of raters were computed both at the level of summated case scores as at the level of item scores. table 3 shows that the correlation of one of the raters (rater 4) with the others falls clearly below all the other intercorrelations. Inspection of the raw data-files showed that the deviations of this rater are caused by a lack of punctuality. Since such a trait could be easily detected in a trial-session it is justified to eliminate this rater from further analysis.

table 3: Intercorrelations between the six raters over all cases

	rater 2	rater 3	rater 4	rater 5	rater 6
rater 1	.90	.89	.80	.90	.91
rater 2		.90	.77	.89	.91
rater 3			.80	.89	.91
rater 4				.78	.80
rater 5					.92

Next, correlations between the raters were computed for each case separately. In table 4 the mean intercorrelations of all six raters and those after exclusion of rater 4 for each case are presented both at the level of case scores and at the level of item scores.

table 4: Mean intercorrelations of sum and item scores per case

case	Sum score 6 raters	Sum score 5 raters	Item score 6 raters	Item score 5 raters
1	.84	.87	.70	.75
2	.66	.76	.66	.75
3	.78	.83	.70	.73
4	.75	.85	.67	.75
5	.74	.78	.67	.70
6	.60	.67	.70	.78
7	.68	.69	.67	.69
8	.76	.87	.63	.70
9	.82	.86	.76	.82
10	.71	.92	.73	.85
11	.84	.89	.73	.77
12	.74	.74	.62	.63
13	.80	.86	.54	.60
14	.34	.33	.40	.44
15	.77	.82	.51	.52
16	.83	.91	.76	.80
17	.70	.79	.72	.72
18	.62	.68	.56	.60
19	.83	.89	.63	.67
20	.79	.83	.80	.82
21	.76	.81	.74	.77
22	.94	.96	.82	.87

The correlations between the raters are as high as could be expected considering the overall reliability. As can be seen, the exclusion of rater 4 results in higher mean intercorrelations, of sum and item scores in all most all instances. At the level of summated case scores only one case (no. 14) shows a low interrater correlation. At the level of item scores the cases 13, 14, 15 and 18 show distinctly lower inter-rater correlations. Reinspection of these cases and the scoring-models by a physician indicated ambiguity of the scoring-model as the probable cause in case 14. A technical problem in the scoring-models explained the low correlations between the item scores in the other three cases. The scoring-models of these cases that were amongst the first to be constructed contained several overlapping items. In the mean time, the scoring-system has been further elaborated and clarified (Van Leeuwen et al, 1987). In the new version symbols are used to distinguish alternative

formulations. Items that are difficult to formulate in exact words are specifically marked. Whether some scoring-models cause problems, can easily be checked with a short trial scoring.
We can estimate the reliability of the scoring of SIMP by a selected nurse, screened for punctuality and scoring with the tested scoring-model, by again computing the Intra Class Correlation with the exclusion of rater 4 and case 14. The result is presented below.

$$ICC = \frac{39.29 - .83}{39.29 + (4 \times .83) + 4(42.94 - .83)/476} = .90$$

The increase of the Intra Class Correlation from .83 to .90 indicates the impact that selection of raters and screening of the scoring models in a trial session could have on the reliability of the ratings.

4.2.4 Discussion

Raters are known to err, thereby increasing measurement error. They are, however, capable of making more complex judgments than machines. Since the influence of rater bias becomes larger with ill defined rating tasks, detailed scoring-models were developed for the scoring of SIMP. With these scoring-models a rater does not have to give a value judgement of the answer, but has only to decide whether or not there is correspondence between the scoring-model and the answer. Knowledge of medical terminology should be sufficient to perform this task. Nurses with some practical experience are likely to possess this knowledge.
In order to find out whether nurses could produce reliable scores with the SIMP-test for medical problem-solving, the scoring of 500 answers by six nurses was analyzed. The rater-reliability was estimated by means of analysis of variance and resulted in an Intra Class Correlation of .83. This indicated that the reliability of scoring by one randomly selected nurse was quite high. The analyses were performed with the Intra Class Correlation because this coefficient yields one estimate of the reliability

for all ratings. However, in the calculations all answers are treated alike, ignoring differences in case mean scores. As a consequence the end result may be somewhat inflated. Nevertheless, it is evident that raters are not the main source of error.

Further exploratory analysis revealed that one rater lacked in punctuality and that one case contained ambiguous items. By testing the raters and screening the scoring-models the reliability of a single random rater could be elevated to .90. The high interrater reliability indicates that the scoring procedures of SIMP may be regarded as objective. Especially, since the nurses are not medical experts they perform their rating task like clerks. An advantage of tests that can be scored automatically, is that scoring by machines is presumably free of error. In fact, however, machines can also make mistakes (for instance by misreading an answer that is marked outside the lines). But, since they are consistent the errors do not show in repeated runs.

The nurses were also highly consistent in their scoring. Just as it would be in the case of machine-scoring, systematic errors (made by all nurses) are not taken into consideration, or in fact contribute to the high reliability.

In order to find out the extend to which the nurses collectively err, part of the answers was scored again by physicians. The results of this extension of the study, with medical experts as raters are reported in the next paragraph.

4.3 Nurses and Physicians Scoring a Test of Medical Problem-solving: the handicap of expertise

4.3.1 Introduction

In the previous paragraph it has been demonstrated that nurses are capable of producing consistent scores with the SIMP scoring-models. Nurses, however, are not themselves experts in solving medical problems. Therefore it is possible that they collectively misinterpreted parts of the scoring-models.
In order to find out the impact of lack of medical expertise on the scoring of SIMP part of the material was scored again by experienced family physicians. The results of the scoring by the physicians and the scoring of the same cases by the nurses are analyzed in the following study.

4.3.2 Methods

In the original study six nurses rated the answers of 25 students on 22 SIMP-cases. Since it was not possible to have all the material rated again by physicians, four cases were selected. One case, which had shown a relatively large disagreement among the nurse raters (case 14, see table 4, p. 50) was specifically included. The other three were chosen, because they were completed by all students. One student who did not complete the first case was removed from the analysis. The remaining 96 answers were scored by two experienced general practitioners.
The overall inter-rater reliability among the six nurses scoring the complete set of 500 answers was remarkably high. Since one case in the present study was specifically selected, because of relatively high disagreement among the raters in the first experiment, somewhat lower reliability coefficients may be expected. Reliability can be approached by analyzing the results a measurement produces when it is replicated (Maxwell and Pilliner, 1968; Mellenbergh, 1977). In the present study

only the facet "raters" is actually replicated. By means of analysis of variance the effects of other replications (other cases and/or other students) can be estimated (Mitchell, 1979).
In this analysis all raters are treated equal, without distinction between the nurse raters and the physician raters. To compare the scores assigned by the two physicians directly with the scores assigned by the nurses t-tests were computed for each case. The scoring was further analyzed at the level of item scores to find out if there were differences in the scoring of specific items.

4.3.3 Results

All answers were scored by all eight raters. The relative contribution of different facets to the total variance is estimated by means of analysis of variance. (see table 5).

table 5: Analysis of variance

source	sum of squares	d.f.	mean square	variance component
mean	34843.66	1	34843.66	45.35
students (S)	394.25	23	17.14	.54
cases (C)	4291.63	3	1430.54	7.34
raters (R)	337.44	7	48.21	.49
SC	1456.09	69	21.10	2.64
SR	231.91	161	1.44	.36
CR	305.24	21	14.54	.56
SCR (Error)	548.80	483	1.14	1.14

The contribution of the raters to the total variance is relatively small. So is, however, the contribution of the Students. Apart from the grand mean the factor Cases represents by far the largest component of variance. This indicates that the reliability of the four cases-test cannot be very high. Indeed generalization to subjects (replication of the measurement with the same cases and raters) gives the following result:

$$R_{xx} = \frac{Vs}{Vs + (Vsc)/4 + Vsr/8 + Vscr/32} = \frac{.54}{.54 + .66 + .05 + .04} .42$$

As might have been expected the test of four cases is too short to produce reliable results. With the Spearman-Brown formula the number of cases necessary to attain a reliability of .80 can be estimated at 22 (employing eight raters).

The reliability of the raters can also be estimated from the variance components by dividing the variance attributed to students and cases and the interaction between them by the variance of all components:

$$p = \frac{Vs + Ve + Vsc}{Vs + Vc + Vsc + Vr + Vsr + Vcr + Vscr} = .80$$

The value of .80 is substantially lower than the .97 that was found for the reliability of the scoring of the complete set of answers by the nurses. Taken into consideration that one case was specifically selected because of relatively low agreement among the nurses this coefficient is fairly good, and shows there are no great differences between the raters. The reliability of the scoring by the nurses and by the physicians can also be estimated separately. When this is done a value of .84 is found for the nurses, and a value of .72 for the physicians. This indicates that the agreement among the nurses is slightly stronger than among the physicians. In order to find out, whether there were systematic differences, between the scoring by the nurses and by the physicians, differences between mean scores of the two groups of raters were tested for significance with a t-test only. On one case a significant difference was found ($p < .01$). Not surprisingly this concerned case 14, which was included because of low agreement among the nurse raters. This difference could very well be accidental, and is not corroborated by other results. On the whole there seems to be no systematic difference between the ratings of the nurses and the physicians.

So far, the analysis have been carried out at the level of composite case scores. This was done because these are the scores that are used for assessment. The case scores can be compared with session scores. Hartmann (1977) points out that the reliability of trial scores is lower than that of session scores. However, there might be some differences between raters, manifest at the level of item scores, that disappear in the adding up of total scores.

The four cases contain 65 items, altogether. For each of the 24 students all eight raters marked a 1, when they figured the item was represented in the answer, and a 0, when it was not. There are several statistical techniques available, for the expression of agreement among raters (Kratochwill and Wetzel, 1977; House et al, 1981). The more sophisticated techniques, like Cohen's kappa, are corrected for chance agreement. Although several methods have been developed to extend kappa for multiple raters, the technique is basically designed for two raters (Fleiss and Cuzick, 1979; Conger, 1980; Schouten, 1986). Agreement above chance level is already demonstrated by the results of the analysis of variance. Therefore, a more uncomplicated solution will do. Agreement of the eight raters on one item, can be expressed as the percentage observed agreement from the maximum possible agreement per item (50% is the absolute minimum, since no more than four raters can disagree with the others on a dichotomous judgement). Table 6 lists the agreement percentages for all 65 items.

table 6: Agreement of eight raters per item

Item	Case 4	Case 5	Case 14	Case 17
1	85	85	66	87
2	85	84	81	82
3	92	89	69	95
4	90	94	88	98
5	89	90	90	88
6	93	92	95	75
7	99	99	96	80
8	90	95	72	94
9	98	90	87	87
10	97	93	84	92
11	82	83	74	95
12	85	93	78	95
13	82	87	96	83
14	92	89		88
15	76	96		86
16	97			96
17	89			
18	81			
19	97			
20	82			
21	91			
mean	89	91	83	89

As might be expected, the over-all agreement at the level of items is high. The average of all cases, except case 14, is above 87.5, indicating agreement of at least 7 of the 8 raters.

Items with a below average agreement can be easily identified. Examination of the scoring-models indicates two different causes for disagreement. In several cases the disagreement can be attributed to the lay-out of the scoring-models. It is possible that the same element of case 14, was scored as item 1 by one rater, and as item 2 or 3 by another.

The other items with relative high disagreement seem to be characterized by an ambiguousness in the formulation, that makes the decision whether that element is represented in the answer less unequivocal. Item 8 of case 14, for instance, asks the raters to decide whether a respondent mentioned "psychic or psycho-somatic problems". This may

be difficult, as it is often implicated in the answer, but not specifically formulated. Another example is item 15 of case 4. On the surface it seems clear enough: "bleeding in the tractus digestivus", but it causes a lot of disagreement, especially between the two physicians. One of them probably demanded a more specific description than the other one. The two physicians disagreed on 20% of the 1540 judgments they both made.

The relatively low over-all agreement between the physicians invalidates their judgement as a criterion for the judgement of the nurses. In those cases where the medical experts disagree, a nurse rater always makes the same decision as that of one of the physicians.

The incidence of complete disagreement was taken as indication of misjudgment by the nurses (all nurses agree on the absence respectively presence of an item, whereas both physicians agree on the contrary). This phenomenon occurred only twelve times (less than 1% of the 1540 judgments). Hence, the lack of medical expertise of the nurses appears to have little consequence.

4.3.4 Discussion

The scoring of open-ended questions is subjected to rater bias. The high agreement among the raters in the above described experiments indicates that the scoring models were successful in enhancing the objectivity of the scoring. First it was demonstrated that the inter-rater reliability among nurses was high.

High agreement among raters, however, does not necessarily mean that the scores are correct. Since nurses are not medical experts, it is possible that they collectively made some mistakes in the interpretation of answers. In order to investigate this supposition the answers on four cases were also scored by two physicians. The overall inter-rater reliability was again high, indicating that there was no notable distinction between different raters. This general impression is not altered by the finding of a significant difference with a t-test between the ratings of the nurses and the physicians on one case.

When analyzed separately, the nurses displayed an even slightly greater consistency among themselves than the physicians did. This could be caused by the fact that the physicians judge the occurrence of an element within the context of their appraisal of the complete answer. Another explanation is that the physicians are less able to restrict themselves to the boring clerical task of scoring the answers without judging.

The nurses appear not to be handicapped by their lack of expert knowledge. It is true that a few instances of complete disagreement between the nurses and the physicians were encountered. The incidence of this phenomenon was, however, too exceptional to attain any significant influence on the scoring.

The nurses are capable of constraining themselves strictly to the scoring instructions. As a result they appear to be able to produce even more reliable scores with the SIMP-test for medical problem-solving than physicians.

Agreement among raters is, however, only one of the facets of test reliability. The reliability of a test can be defined as the replicability of the results of a measure (Mellenbergh, 1977). The analysis of variance shows that the component cases and the interaction between cases and students explain a large proportion of the variance. This means that the attribute that is measured is not homogeneously distributed over cases. As a consequence, a substantial number of cases is necessary for a stable measurement. The reliability of the four-case test was .42. With the Spearman-Brown formula for test length the number of cases necessary to attain an acceptable reliability of at least .80 can be estimated at 22.

In order to make a test more reliable the administering of extra cases is more effective than the addition of raters. In the next paragraph a generalizability analysis is presented of the effects of the number of raters and cases on the test reliability.

4.4 Generalizability of SIMP-scores

The preceding studies focus on the agreement among raters. Inter rater reliability is, however, only one of the facets of the reliability of test-scores. As mentioned before generalizability theory provides a framework that takes several facets into account at the same time. Since generalizability analysis is based on analysis of variance, the design must be fully crossed in order to allow generalization to all facets (i.e. each student must have completed all cases, and all answers must be rated by all judges). Because several students did not complete all cases the data matrix of the study with the nurse raters contained a substantial number of missing values. Hence, a generalizability analyses with all facets could not be performed. Since inter-rater reliability was of primary interest, the facets cases and students were collapsed into "answers", resulting in a fully crossed matrix of raters x answers.

In the selection of the four cases for the second study the problem of missing values was averted. After the omission of one student a fully crossed matrix with eight raters, four cases and 24 students remained. With these data a generalizability-coefficient of .42 was computed, indicating that a substantially longer test will be needed to ensure reliable scores. The generalizability analysis can be extended to estimate the effects of increasing test length and manipulating the number of raters with a D (decision)-study. The results of increasing the number of cases in steps of four, up to 32 (more than three hours of testing time) with one and with two raters are displayed in table 7.

table 7: D-study, based on four cases and eight raters

number of cases	generalizability coefficient	
	one rater	two raters
4	.29	.35
8	.39	.48
12	.44	.55
16	.48	.59
20	.50	.61
24	.52	.63
28	.53	.65
32	.53	.66

As can be seen the reliability of the test remains rather low. If one rater is employed a generalizability coefficient of .53 is reached for 32 cases. A second rater increases the reliability to .66 which however is still not sufficient. Both further increasing of test length and increasing the number of raters are inefficient in terms of time and cost. The four cases on which these calculations are based, however, cannot be regarded as representative for SIMP-cases. One of the selected cases even had shown problems with the reliability of the scoring before. The study with the nurse-raters contains a larger sample of cases, but the high proportion of missing values precluded a full scale generalizability analyses. Several methods to deal with this problem are available. Incomplete records can be deleted, or sophisticated techniques can be used to estimate missing values (Frane, 1979).

However, the removal of all incomplete records would leave almost no data to analyze, and it seemed preferable not to use estimates. Therefore a combination was made of deleting three records and omitting eight cases. The resulting data matrix contains 22 students and 14 cases, fully crossed with the six raters. The results of the D-study based on this data matrix are displayed in table 8.

table 8: D-study with fourteen cases and six nurse raters

number of cases	generalizability coefficient	
	one rater	two raters
10	.42	.42
14	.51	.54
18	.57	.60
22	.62	.65
26	.66	.69
30	.69	.72

The resulting generalizability coefficients are slightly better, though still not sufficient. With 30 cases and two raters a generalizability coefficient of .72 is reached. This disappointing reliability is explained by the rather low proportion of student variance, compared to the relatively high case variance and student by case interaction. In order to discriminate reliably within this homogenous group of students a long test is necessary. The reliability can be increased more efficient by adding cases than by employing more raters.

However, since several students and cases were removed from the analysis the data are not altogether representative. Hence, definite conclusion regarding the desirable test length cannot be drawn.

5 THE VALIDITY OF SIMULATION OF INITIAL MEDICAL PROBLEM-SOLVING (SIMP) AS A MEASURE OF CLINICAL COMPETENCE[1]

5.1 Introduction

The most serious problem in the assessment of medical problem-solving is doubt on the validity of measurements. A measure of medical problem-solving is intended to reflect the ability to respond adequately to problems in medical practice. Validity stands for the degree to which this intention is realized. Criticism on the validity of a measure directly affects any conclusions to be drawn from it.

In research with Patient Management Problems (PMP) construct validity was demonstrated by a positive relationship with educational level (Schumacher et al, 1975; Robinson, 1977). Concurrent validity was established by the repeated demonstration of significant (though rather low) correlations with measures of other competencies, such as multiple-choice measures of knowledge (Schumacher, 1973; McGuire, 1973; Mura et al, 1976). Comparable results were reported for the open-ended Modified Essay Question (MEQ) (Feletti, 1980; Rabinowitz, 1987).

Other observations, however, raise serious doubt on the validity of measurement of medical problem-solving. A common feature of all approaches is that the scoring is based on a sum of weights assigned to individual options selected by the candidate. This approach to scoring places heavy emphasis on thoroughness of data gathering and has been shown to favor inexperienced problem-solvers (Newble et al, 1982). More problematic is the repeated observation of a low correlation of scores

[1] Adapted from: De Graaff, E., Post, G.J. and Drop, M.J. (1987). Validation of a new measure of clinical problem-solving, Medical Education, 21, 213-218.

across different problems; a phenomenon which has been labelled "content specificity" by some authors (McGuire, 1976; Elstein et al, 1978) and which raises serious questions about the ability of these measures to assess a general "problem-solving ability".

There are several possible explanations for the problems in the assessment of problem-solving skills. One insight is derived from basic research into the nature of clinical problem-solving (Elstein et al, 1978; Harasym et al, 1980; Barrows et al, 1982), where it was found that the generation of early diagnostic hypotheses played a central role in successful problem-solving. Furthermore, these studies have shown that thoroughness of data gathering was uncorrelated across problems and unrelated to successful diagnostic and management outcomes.

More recent research has shown that clinical problem-solving is more automatic and less analytical and logical than might have been expected (Norman, 1985-a; Groen and Patel, 1985). If this is the case, then the first brief period of time after the students encounter with a clinical problem might be critical to the assessment of the ability to solve medical problems.

A more technical assessment problem is the effect of "cueing" of the given options, which has been shown to contribute to the low validity of PMPs (Goran et al, 1973; Feightner and Norman, 1976; Page and Fielding, 1980; Norman and Feightner, 1981). An open-ended format seems an appropriate way to minimalize this effect.

SIMP combines emphasis on the first impression after confrontation with a patient problem, with an open-ended question format. The concurrent validity of SIMP was first investigated by De Graaff and Galesloot (1982), who compared SIMP to other measures assessing the same or similar competencies. The results of two SIMP-tests were correlated with a global competence judgement by their tutor (a family physician), a rating of performance with a simulated patient (reported in Bouhuijs, 1983) and scores on two Maastricht progress tests. The results of these correlations are shown in table 9.

table 9: Correlations of two SIMP-tests with other competency measures

	Tutor rating	Simulated patient	Progress test 1	Progress test 2
SIMP 1	.14	.43	.10	.25
SIMP 2	.27	.74	-.16	-.06

The correlations indicate no substantial relation with the knowledge test. In contrast the correlations between the simulated-patient and the SIMP-tests are relatively high. Therefore it would appear that scores on the written SIMP-test were related to a concurrent measure of clinical performance and not to the assessment of factual knowledge. This result was considered as promising with respect to the measurement goal of SIMP. In order to confirm and extend these results a second study was conducted in 1982.

5.2 Methods

From a review of the literature on measurement of medical problem-solving it was concluded that the Patient Management Problem (PMP) was the most well known and most widely used concurrent measure of SIMP (Post et al, 1985). Since these are also paper-and-pencil simulations of medical problem-solving these measures relate to performance in practice in a similar way as SIMP. An experiment was designed with two PMPs, two Simulated Patent Encounters (SPE) and six SIMP-cases. The simulated patients were trained at the Skills Laboratory of the University of Limburg. One simulated patient presented respiratory problems, the other lower back pain complaints. The two PMPs were developed to match these same problems. In the SIMP-test each subject was covered by three cases.

Subjects in the study were 29 recently graduated "basic physicians" of the University of Limburg (19 male and 10 female). All participants had graduated less than one year ago, and 13 had yet none or almost no

experience in practice after graduation. Median duration of study in the group was 5.9 years, which is shorter than usual in the Netherlands, but about equal to other cohorts of Maastricht graduates (Post et al, 1988).
The experiment was conducted in the Behavioral Laboratory, which made it possible to videotape the Simulated Patient Encounters and to observe the performance of the candidates from behind a one-way screen. Since only two simulated patient encounters could be taped at the same time the tests were administered in a sequential scheme. First the SIMP-cases, next half of the graduates completed the PMP followed by the Simulated Patient Encounter, the other half vice versa. The total test session, including a one hour break, took about four and a half hours.
The SIMP-cases were rated with the scoring-models by four experienced judges, the PMPs were scored by a single rater according to the instruction and the Simulated Patient Encounters were scored from tape by 3 to 4 judges (two judges scored all cases and subjects). The raters used a detailed checklist (SPE 1) and also gave a global judgement of performance (SPE 2). In all cases where multiple raters were employed, mean scores over raters were computed. Scores on the Progress Test (PT) as described earlier were added as a comparison with medical knowledge.

5.3 Results

The reliability of the SIMP-scores was investigated with generalizability theory (Cronbach et al, 1972; Mitchell, 1979). The results of the analysis of variance with the general mixed model and two repeated measures factors are shown in table 10.

table 10: Analysis of variance on SIMP

source of variance	sum of squares	d.f.	mean square	variance component	percentage of total variance
graduates (g)	931.94	26	35.84	1.10	21.15
raters (r)	2.42	3	.81	-.03*	.00
cases (c)	620.31	5	124.06	1.03	19.81
r x g	101.45	78	1.30	.08	1.54
r x c	74.47	15	4.96	.15	2.89
g x c	1160.31	130	8.93	2.03	39.04
r x g x c + error	317.40	390	.81	.81	15.58

* a negative component is assumed to be zero.

The zero-variance of the raters can be interpreted as indicating there was no systematic bias among raters. The main sources of variance were the graduates, the cases and the interaction among them. The large proportion of variance that is attributed to the cases suggest that the ability that is measured, is not homogeneously distributed over cases. As this was not expected this result need not be discouraging. In order to be reliable a test must consist of a substantial number of cases. The generalizability coefficient across cases yields .74 estimating the reliability of the 6-item test).

$$p2 = \frac{Vg}{Vg + (Vg*c)/6 + (Vg*r)/4 + (Vg*c*r+e)/24}$$

$$= \frac{1.10}{1.10 + .34 + .02 + 03} = .74$$

V = variance-component; g = graduates; c = case; r = rater

Generalizability-coefficients are usually lower than the usual alpha reliability coefficient (Mitchell, 1979), therefore this result is remarkably good. Based on the analyses of variance the effects of changes in the

number of raters and cases can be estimated. Since raters evidently added little to the variance, computations were made only for one and for two raters scoring all answers. The effect of increasing the number of cases in steps of four is presented in table 11.

table 11: D-study, number of raters and cases

number of cases	generalizability coefficient one rater	two raters
6	.65	.71
10	.75	.79
14	.79	.84
18	.82	.86
22	.84	.88

From this table it follows that a SIMP-test of about 15 cases, scored by one rater, is sufficient to attain a reliability of .80. With two raters scoring all answers on 11 cases will do.

In order to compare the results of SIMP with the PMP and the Simulated Patient Encounter a summed-up total score for each instrument was calculated. The results of the correlations are presented in table 12.

table 12: Correlations between different instruments for measuring medical competence.

	SIMP	PMP	SPE 1	SPE 2	PT
SIMP		-.07	.19	.38*	-.25
PMP			.00	-.03	-.42*
SPE 1				.45*	-.13
SPE 2					.05

* Significant, $p < 0.01$.

The results replicate and extend the findings of the previous study. One of the correlations between SIMP and the two scores based on the

patient encounter was significant positive, the other slightly positive (.38 and .19). In contrast the correlations with the knowledge test were not significant, tending to be negative or zero. Of interest is that the PMP-score showed no positive correlation with any of the other sub tests.

5.4 Discussion

The generalizability analyses of SIMP in this study displayed remarkably good reliability figures. As a result the testing time necessary to obtain reliable scores is relatively short. A test of 15 SIMP cases takes about two and a half hours, with 11 cases two hours would be enough. Such a gain in testing time is however more than compensated by the loss in rater time. One rater takes about 45 minutes for the scoring 15 answers of one candidate (3 minutes per case, see chapter 4.2), whereas two raters together would spent 66 minutes on the scoring of 11 cases.
The discrepancy between the results of the generalizability analysis in this study and those reported in the previous chapter may be explained by the methodological problems in that study. The non random elimination of cases and subjects affects the generalizability of the data. In both studies, however, the number of subjects is rather small. Therefore the difference between the generalizability coefficients may be interpreted as indication of the low reliability of the estimates. This means that the actual test reliability could be either higher or lower than the reported coefficients.
An alternative explanation is constituted by the supposition of content "specificity". The difference between the generalizability coefficients could reflect a difference in content of the respective SIMP-tests. The low generalizability coefficient was found in a study with 22 cases (16 left for the analysis) covering a variety of subjects. The high generalizability coefficient is based on an experiment with six cases, related to no more than two problem fields. This suggests the possibility that the high generalizability only holds true for generalization to a limited domain.
Since all measures in this study cover the same domain the correlational analysis of the construct validity is not affected. The correlations be-

tween the different measurements of medical competence suggest the existence of two underlying factors - medical knowledge, as assessed by the Progress Test and clinical competence, common to both the SIMP and the Simulated Patient Encounter. It is encouraging that SIMP appears to relate relatively strongest to the Simulated Patient Encounter, since this indicates that it is assessing an ability more closely related to clinical performance rather than to knowledge. These results do not hold true for the PMP, which shared no common variance with any of the remaining tests.

Nevertheless, the correlations across tests are not high. The hypothesis described before as "content specificity" suggests that the content of a case influences the score and thereby reduces the correlation between different cases. In a recent study (Norman, 1985-a) evidence was found that "content-specificity" is not sufficient as an explanation of a weak correlation between cases. An alternative explanation is that each case should in fact be considered as a single item in a test, therefore a moderate correlation across items is to be anticipated. Elaboration of this assumption produced some further confirmation. Both the global judgement of the Simulated Patient Encounter (SPE 2) and the score on SIMP reflect more than just the performance with respect to a specific content domain. The correlation between these two measures is substantially higher than between any of the other measures. (Naturally apart from the correlation between the two ratings of the SPE).

Either on the basis of this last explanation or on the basis of the assumption of content-specificity the conclusion, that a test for medical competence must consist of sufficient cases (items), is clear. In this respect SIMP has the advantage over the PMP and the SPE that the administering of one case takes considerably less time than either alternative.

Summarized altogether the results of this study provide support for the construct validity of SIMP. With respect to the content validity evidence was found suggesting that the generalizability of medical problem-solving is to some extent domain specific. If cases should be regarded as items in a test this would imply that a measure of medical problem-solving

should consist out of 10 to 15 cases within each content domain. However, further research is necessary to investigate the impact of content sampling on the generalizability of SIMP.

6 DISCUSSION AND CONCLUSIONS

6.1 The construct medical problem-solving

Simulation of Initial Medical Problem-solving has been developed as an operationalization of the construct medical problem-solving, to be used for the assessment of medical students. Ideally, such an operationalization should be based on a sound theoretical framework and a clear definition of the construct. At the beginning of instrument development, however, there was no generally accepted theory available. A search of the literature even indicated that the construct medical problem-solving has several different connotations. The general notion seemed to be that the ability to solve medical problems reflects some kind of mental strategy or cognitive reasoning skill. These vague notions did not appear to be a sound basis for operationalization. Later research on thought processes indeed casts serious doubt on the existence of such a skill apart from normal adult thinking (Gale, 1981). Neufeld et al (1981) found that medical students used the same problem-solving process throughout medical school. They concluded that if it is a skill at all, it is not learned in medical school. In a recent review of more than ten years of research on medical problem-solving Norman (1988) points out that, even if these (problem-solving) skills do exist, they do not go far in explaining the acquisition of expertise. And the training of medical experts is precisely what is intended when medical problem-solving abilities are emphasized as an important goal of medical education.

The working definition of medical problem-solving that was formulated as a foundation of instrument development focussed on medical actions. It

was reasoned that the competence of a physician in practice situations depends on the ability to act directly. The reasoning behind these actions does not have to be fully conscious. It may even be partly a matter of routine.

Extensive research with simulated patients suggests that physicians formulate tentative diagnoses or working hypothesis at an early phase of the problem-solving process (Elstein et al, 1978). The rational logic implied in the labelling as "hypothetico-deductive method", for the ensuing process of gathering additional data until a final diagnoses is reached does, however, not appear to fit the normal behavior of a physician in practice. Rather than listing possible explanations and considering tests to confirm or refute them, a physician who encounters a patient immediately starts asking questions or examining the patient. Recent research of the cognitive processes of physicians indicates that the differences between medical experts and novices in generating hypotheses can be explained by differences in knowledge structure (Schmidt et al, 1988; Boshuizen et al, 1988). Experience with similar cases organizes knowledge and provides direct access to appropriate actions.

SIMP offers case descriptions including ambiguous context information, followed by an open-ended question. The original idea was the ill-structured problems, that are generally encountered in medical practice call for an open-ended question, as there are many different correct responses to those problems. The correctness of this view was corroborated by a review of Frederiksen (1984). In general there seemed to be little evidence in favor of open-ended questions, but with respect to a task of formulating possible explanations for a given phenomenon, evidence was found that the two formats (open-ended and objective) do not measure the same construct. This task appears to have similar characteristics as the operationalization of the ability to solve medical problems by means of SIMP.

Both the decision to focus on clinical actions instead of the reasoning process as well as the choice of the open-ended test format appear to be supported by research findings. The relationship of SIMP measure-

ments with the recent insights on the cognitive structure of physicians need, however, to be investigated further.

6.2 Objectivity

Open-ended questions are put forward as an important characteristic of SIMP. The disadvantage of the open-ended question format, however, effective as it may be in simulating practice conditions and in preventing queing, is that the answers must be scored by human raters, a procedure that exposes the measure to subjective rater bias. Although rater bias is only one of the sources of error that influence measurement reliability, objectivity is usually regarded as a necessary condition for any measurement procedure.

Because objectivity could not be attained by mechanic procedures, the objectivity of the scoring procedure of SIMP was enhanced by the development of scoring-models. Thereby the task of the raters is reduced to recognizing correspondence between elements of the answer and items of the scoring-model. Since no value judgement is demanded, the raters may operate as clerks, following a uniform instruction.

The reliability of this scoring procedure was investigated in two experiments. First a set of answers was scored by nurses, on the assumption that their knowledge of medical terminology was sufficient to perform this task. Although one of the nurses performed less well than the others, and one of the scoring-models was found to cause scoring errors, the overall agreement among the nurses was high. This result can be interpreted as support for the objectivity of the scoring procedure.

Though nurses may agree on judgments with respect to medical problem-solving their scores could be biassed by their inability to judge medical competence: high agreement might result from a lack of understanding common to all nurses. In a second study this possibility was investigated through an additional scoring by two physicians. The overall agreement between the raters, including nurses and physicians, was high. Detailed analyses of the scoring, however, revealed some interesting differences between the two groups of raters. In about one percent

of the scores, all nurses agreed on a rating that differed from the ratings of both physicians. Though such deviations may indeed be regarded as indication of a lack of insight from the part of the nurses, the impact on the total scoring was small.

More important was the finding that the agreement among the physicians overall ratings, was lower than that among the nurses. Apparently, physicians are less suitable to be employed as clerks in the scoring procedure. One explanation for this phenomenon might be that it is difficult for physicians to disregard their own (global) interpretation of the right way to solve the problem. Another explanation is that the physicians are unable to restrict themselves to the performing of simple clerical tasks during a longer period of time.

Summarizing, it may be concluded that on the whole the scoring-models are on the whole effective in controlling subjective influence on the scoring, although careful surveyance over the raters remains necessary to ensure their adherence to the prescribed procedures. In this respect, expertise in medical problem-solving seems to be a handicap rather than an advantage for the raters.

6.3 Instrumental utility

The basic question with each new instrument is: does it add something worth while to the already existing variety of measures? Usefulness of a test is a function of its measurement properties, like reliability, validity and of economic factors. Assessment of medical problem-solving seems to suffer from a conflict between the demands of reliability and the demands of validity. A reliable test should be standardized, the contents homogeneous and the scoring objective. Validity - the degree to which a test measures what it is intended to measure - is served by close resemblance to actual performance in practice. In real-life practice, however, it is very hard to control those measurement conditions.

The intricate relationship between reliability and validity is demonstrated by the lack of consistency among cases (Norman et al, 1985). When cases are regarded as items in a test, low consistency among cases

results in low test reliability. Indirectly the validity is also affected since the maximum validity that can be attained is equal to the square of the reliability (Guilford, 1954). Hence with respect to both reliability and validity a test for the assessment of medical problem-solving should consist of a substantial number of different cases.
At this point the economic factor enters the picture. There is a limit to acceptable testing time. Therefore, it is not always possible to include as many cases as would be desirable. The question is whether a reasonable balance can be found: sufficiently reliable test results to allow for valid measurement, within an acceptable testing time.

6.3.1 Reliability

There are several factors that simultaneously influence measurement reliability. In principle, everything that might cause differences between replications of a measurement affects the reliability. Ideally, repeated measures of the same trait would result in the same outcome. In practice, a reliability coefficient of .80 is usually taken as an acceptable criterion.
With generalizability analysis the effects of all facets can be estimated at the same time. Due to incompleteness of the original data set, generalizability coefficients could only be determined for part of the data of the scoring reliability studies, reported in chapter 4. In these sub-sets reliability proved to be rather disappointing. The largest proportion of variance was attributed to cases and the interaction between cases and persons; the persons variance was relatively small. As a consequence, a test of more than 30 cases would be needed to reach a reliability of .80. That would imply a testing time of more than four hours, and an enormous task for the raters.
The generalizability study reported in chapter 5 displayed high rater-reliability. The variance attributed to the raters was even estimated zero. The generalizability coefficient of .74, represents the reliability of a six case test, scored by four independent raters, in a population of graduates. In order to achieve a reliability of at least .80, the number of cases

should be increased to about 15 with one rater and to 11 with two raters.

There are several explanations for the remarkable discrepancy between the generalizability coefficients in the different studies. First, the number of persons in both studies was rather low, which results in unreliable estimates of variance components. Second, as a consequence of the large proportion of missing values, a number of cases and persons were eliminated from the data set of the study reported in chapter 4. The representativeness of the remaining data is affected. Finally, the more homogenous content of the SIMP-test in the study reported in chapter 5, may have inflated the generalizability coefficient. The six cases in this experiment involved only two complaints, (three cases each). Therefore, it is not justified to generalize from these data to a general problem-solving ability.

The available data seem to warrant the conclusion that it is possible to contrive a sufficiently reliable SIMP-test within a reasonable testing time and scored by one rater, at least within a limited domain. However, further research with respect to the effects of content sampling is necessary. Also replication with larger groups of subjects will be needed to establish more reliable generalizability coefficients.

6.3.2 Validity

Validity is characterized by the question: are we measuring what we think we are measuring? This question accentuates both the importance of validity, and the difficulty of assessing it. An instrument can only be valuable, if it accomplishes its purpose. So the first requirements are to define exactly the intent of the measurement. As SIMP was constructed for the assessment of medical problem-solving, predictive validity could be established by comparing SIMP-scores with ratings of performance in practice. However, as stated in chapter 2, a gold standard of practice performance does not exist. In an article on validity of licensure examinations, Kane (1982) states that the emphasis on predictive validity is not justified. He argues that examination scores should be interpreted in

terms of specific abilities, that are critical for professional practice. The validity of a test designed to measure such an ability would depend on the importance of that ability for practice and on the quality of the measurement. The first question is whether medical problem-solving represents such a critical ability for medical education. Two decades of research suggest that the importance of this construct is beyond any doubt. The results of this research, however, indicate that it has many un-skill-like characteristics (Norman, 1988). For instance, problem-solving skills do not seem to change during medical education (Neufeld et al, 1981). And differences in knowledge related to the level of education appear to explain differences in problem-solving skills (Maatsch et al, 1982). Since, the working definition of medical problem-solving, presented in chapter 2, emphasizes the ability to respond with adequate actions to situations in medical practice, instead of the mental process of problem-solving, these criticisms do not need to apply to SIMP. Medical experts and students do evidently differ in their ability to solve medical problems. The question remains whether SIMP is a valid measure for this ability.

In order to investigate the concurrent validity of SIMP correlations with other measures were determined. The most important result was a moderately positive correlation of SIMP with global rating of a simulated patient encounter. This was interpreted as support for the validity of SIMP as operationalization of problem-solving in medical practice. Since PMPs correlated negatively with all other measures in this study, the negative correlation with SIMP, was interpreted as confirmation the lack of validity of the PMP. The lack of significant correlation with knowledge tests was also interpreted in favor of SIMP, since high correlations would implicate that SIMP measures the same attribute as the knowledge tests. Support for the validity of a measure can also be found in its discriminative power. De Graaff and Galesloot (1981) demonstrated that SIMP has the ability to discriminate between students of different levels of experience. They report a distinct growth curve for two SIMP-cases that were on a voluntary basis added to a Progress Test session. However, the impact of such support is limited. As Swanson et al (1987) observe,

it would be a poor measure indeed, that does not differentiate between obvious levels of competence. Nevertheless, inability to discriminate would clearly have invalidated the measure.

Some more information bearing on the validity of SIMP is provided by investigators, who have applied the instrument during the past years in other research projects. The relation between SIMP and ratings of performance with a simulated patient, was investigated by Crijnen et al, (1987). In a validation study on a newly developed instrument for the assessment of medical interviewing abilities (the Maastricht History-taking and Advice Checklist; MAAS), high correlations were found between components of the SIMP score and MAAS sub-scales, relating to corresponding subjects.

The relation between SIMP-scores and medical knowledge, was further investigated by Van Leeuwen (1987). In a comparison between a ten case SIMP test and matched knowledge test no significant correlation was found. This result, however, may be contaminated by the low discriminatory power of the knowledge test. The reliability of the SIMP test was again satisfactory. Van Leeuwen further reports manifest satisfaction of the respondents with SIMP. They appreciated SIMP as an extension to the knowledge test.

A very basic question regards the validity of the construct: does the operationalization cover the original concept-as-intended? The operationalization of medical problem-solving by SIMP is limited to written simulation of general practice situations, and concentrates on the beginning phase of the encounter. The relation between SIMP assessment and real performance in general practice was investigated by Rethans and Van Boven (1987). They did send under-cover simulated patients to the practice of 48 family physicians. The performance of the physicians with the simulated patient was compared with the results on a similar SIMP-case. The overall scores of both measurements showed no significant differences. Detailed analysis, however, revealed that the physicians performed several relevant acts with the simulated patient that they did not write down on the paper simulation. Furthermore, they wrote down several unnecessary and superfluous actions that they did not perform in

practice. This suggests their real performance was better than indicated by the answers on the written SIMP test. Partly this phenomenon might be caused by the fact that respondents react differently when they are aware of being judged. Such a limitation of the validity of any test with respect to prediction of performance in practice is inevitable. Taking into consideration that this limitation is inherent in any examination, there seems to be no reason to suspect that this phenomenon particularly belongs to SIMP.

Summarizing, the validity of SIMP is supported by positive correlations with performance ratings on a simulated patient encounter. Further the instrument displays discriminative power, and differentiates from assessment of pure knowledge. So far all available data suggest that SIMP might be a valid measure of medical problem-solving.

Several questions, however, still remain to be answered. For instance, the extent to which SIMP-scores depend on specific medical knowledge must be further explored in order to find out what SIMP adds to a good knowledge test. The cognitive approach of medical problem-solving which emphasizes "structure of medical knowledge" (Boshuizen, 1989), seems to offer a more compatible theoretical framework than the hypothetico-deductive model. The relationship between SIMP and measures derived from this procedure need to be investigated.

6.3.3 Efficiency

The efficiency of a measure is defined as an optimal balance between cost and quality. Cost is expressed in terms of time spent on test construction, testing time, scoring time and material costs like special devices that are needed.

As a paper-and-pencil test SIMP is less expensive than Simulated Patient Encounters, which require trained simulated patients, video equipment, raters, and a lot of testing time. The scoring of answers on the open-ended questions also necessitates the employment of raters. Notwithstanding the high inter-rater reliability, which indicates that sufficient reliability can be obtained with a single rater, the cost of

scoring increases linear with an increase of candidates. The undisputed advantage of tests that can be scored automatically is, that it makes almost no difference whether there are 10, 100 or 1000 candidates.
Therefore the slight advantage of SIMP over PMPs in construction time, turns into a disadvantage when large numbers of students are involved (disregarding the difference in testing time, which is in the favor of SIMP).
With respect to quality, there is evidence indicating that SIMP may add relevant information to tests of factual knowledge. Although the actual correlations were low, the consistent relationship with Simulated Patient Encounters suggest SIMP truly indeed might succeed in measuring the ability to solve medical problems in practice. Both from the perspective of validity as from that of reliability a substantial number of cases within one test is required.
The results of the generalizability analyses, that were carried out to assess reliability, are somewhat contradictory. Apart from the methodological problems, which may explain at least partly the low generalizability coefficients in the studies reported in chapter 4, it seems likely that the sampling of test content strongly affects reliability. Within a limited domain a SIMP-test of 15 cases is estimated to be sufficiently reliable. Such a test can be administered in about two and a half hours.
The considerations which decide whether there is a reasonable balance between the importance of the extra information provided by SIMP and the costs of gaining that information depend on the situation in which the test is applied.

6.4 Applications of SIMP

Although further research on the reliability and validity remains necessary, the research reported in this thesis suggests SIMP can be utilized as a measure of medical problem-solving abilities. In research on medical competence SIMP could be employed as criterion or concurrent measure, as was already done in several studies cited above. With respect to the validity of the construct medical problem-solving further

investigation of the relation of SIMP scores with performance in practice would be of special interest. Such research could provide information on the factors determining quality of performance and their representation in the measure.

Also important is further exploration of the relationship between medical problem-solving abilities and factual medical knowledge. Both correlations of SIMP scores with corresponding knowledge tests and comparison of SIMP results with assessment of the structure of medical knowledge could provide more information regarding the question how and to which degree medical knowledge constitutes a condition for the ability to solve medical problems.

The development of SIMP was however enhanced with the purpose of extending the range of measurement of the Maastricht Progress Test. In the research reported in this thesis SIMP has at least been demonstrated to have the potential of fulfilling that purpose. Yet, SIMP has not been adopted by the medical faculty as an extension of the Progress Test. Nor is it very likely that it will be applied in that manner in the future.

The reason for this refutation is that the cost of implementing SIMP alongside the Progress Test would be enormous. The Progress Test is administered four times a year to about 900 students from freshmen to nearly graduated physicians. Supposing a SIMP-test of 30 cases that reliably covers the total domain of medicine could be constructed each time, not only five hours would be added to the testing time, but the faculty would also have to supply about 135 hours of rating time, almost a month work for one person.

The evidence of cognitive research on medical problem-solving cited earlier, suggests that such a large claim on faculty means is not justified. If the ability to solve medical problems depends on the structure of medical knowledge, the testing of this ability would only make sense with subjects that posses sufficient knowledge. One alternative for the application of SIMP as part of the system of progress testing would be to define a minimum score level on the knowledge test as admission criterion for a problem-solving part to be taken later on. That way

students without enough factual knowledge could be excluded. However, even then the logistic problems will still be large.

The efforts of regularly testing large groups of students (let's say more than 100) will never easily counterbalance the profits. For such a purpose it would be much more profitable to try and develop a test that can be scored mechanically. In this respect the high agreement among the nurses in rating with SIMP-answers constitutes a promise. When it is possible to obtain reliable scores from non-expert raters, it may very well be possible to develop computer programs that can perform this function in the near future.

Another alternative would be the development of an objective scoring-system which prevents cueing by offering a large surplus of options. In that case, however, the advantage of the open-ended question in representing of the situation of a physician in practice would be lost. Therefore the validity of such a measure should be thoroughly investigated, before it is put to use.

More in line with the reported results of research with SIMP would be to look for possibilities of application within limited domains and relatively small groups of students. For instance SIMP could easily be adapted for evaluation in the clinical clerkships. Analogous to the tests that were used for formative evaluation in the family medicine clerkship (see chapter 4), SIMP-like measures could be developed for other clerkships, or in fact for any health care profession. For instance SIMP could be part of the assessment in the post academic family medicine training program, and SIMP like measures could be developed for the assessment of professional nursing abilities. Exemplary case material could be readily supplied by experienced professionals and subsequently adapted for assessment.

Application of SIMP for assessment within the clerkships has distinct advantages. The assessment of students during the clerkships focusses on their ability to solve practical medical problems and guide them on their way to becoming trustworthy physicians. As a relatively simple paper-and-pencil measure SIMP could be useful in providing standardized feed-back to students.

Apart from application within the summative assessment system the feed-back from SIMP can also be utilized as a means for self-evaluation (Galesloot and De Graaff, 1981).

6.5 Conclusions

A clear conclusion that can be draw from the research in this thesis is one that was already evident from the first literature search: the construct medical problem-solving is hard to measure. It is difficult to formulate an indisputable definition to begin with. But the biggest problem is to set standards for a domain as large and diverse as the problem-solving behavior of physicians in medical practice.
In the construction of SIMP this problem is avoided rather than solved. No attempts have been made to define a "gold standard" of ideal problem-solving. Instead the method relies on experts who implicitly define a standard by selecting cases and constructing scoring-models.
The lacking of a universal gold standard was also an important argument to choose for the open-ended question format. When experts do not agree on one single best approach of a problem, an assessment method should not force students to believe that such an approach exits. As a consequence rating time is added as a cost factor, which is hard to control when the number of students grows larger.
Actually, cost is an important factor in all educational assessment procedures. Any assessment method requires time of staff and students. Although students can learn from tests, and these can be helpful in directing study activities, assessment can also act as a barrier, hindering students in the progress of their study (De Groot, 1973).
Educational assessment may serve several different purposes at the same time. Wesdorp (1979), distinguishes three basic categories of assessment functions, bearing on: A. student careers, B. management, and C. educational content and form. These functions are not mutually exclusive. Assessment results, collected for decisions about students progress, may also serve management purposes. In the long run curriculum adjustment may result. More directly, assessment causes stu-

dents to adapt their learning strategies, a phenomenon labelled "the didactic function of assessment" (De Koning, 1977).

Within the context of problem-based learning the didactic function of assessment is particularly important. The system of progress testing aims at operationalizing the end level of the medical school. As examinations fulfill a function of directing the study activities of students, it is essential that important educational goals are represented in the examinations. Otherwise, students might be tempted to concentrate on those elements of the study they know for certain to be part of the assessment. Since there seem to be general agreement on the importance of medical problem-solving, or the ability to perform in real practice as a final goal for medical education (see chapter 1), this ability should be represented in the examinations of the medical school.

SUMMARY

Measurement of Medical Problem-solving

Medical problem-solving is often designated as an important general goal of medical education. Therefore, measurement of medical problem-solving abilities should be represented in the assessment of medical students. The general objective of medical education is to train physicians, who are capable of functioning in several different positions within our modern health care system.

At the Rijksuniversiteit Limburg a special instrument has been developed for the measurement of increase of medical knowledge: the Maastricht Progress Test. As extension of the Progress Test, a measure was needed for the assessment of medical problem-solving abilities.

Research with existing instruments, like the Patient Management Problem (PMP) and the Modified Essay Question (MEQ), indicated these measures suffered from several weaknesses. The results of PMPs, for instance, were found to be biassed by a cueing effect of the optional answers. Furthermore, the consistent finding of a low correlation among cases casted doubt on the validity of the measures. Therefore it was decided to initiate a project for the development of a new measure for the assessment of medical problem-solving.

Simulation of Initial Medical Problem-solving

Basic research into the nature of medical problem-solving has indicated that the ability to formulate initial hypotheses, directly in the beginning of the contact has crucial impact on the rest of the encounter. Therefore, it seemed sensible to focus instrument development on the initial moments of the patient-physician encounter. Furthermore, the open-ended question format seemed a logical means to avoid the cueing-effect of optional answers.

Based on these principles a measure was developed, called Simulation of Initial Medical Problem-solving. The instrument consists of short case histories, followed by one single open-ended question: "What would you do as a physician in this situation?" This question puts almost no constraints on the respondents. The responses are only limited by the information from the presented cases. Feed-back is given after the test is completed. The answering of one question takes five to ten minutes. Hence, a two hour test may contain ten to twenty different cases.

Objectivity of measurement

A well known disadvantage of open-ended questions, is the subjective influence of raters. In order to reduce the effects of rater bias, scoring-models have been developed. These scoring-models take the format of checklists, describing elements of the correct answer. To facilitate scoring, the scoring-models are organized according to the SOAP-system for medical audit (Weed, 1969). Respondents, however, remain free to chose their own answer structure.

With these scoring-models, raters are not expected to judge the value of an answer. All they have to do, is mark the similarities between a respondents answer and the scoring-model. This kind of judgement does not demand expert medical knowledge. It is sufficient, if a rater understand medical terminology, in order to recognize synonymous expressions. Since, nurses, by training and experience, fulfill these requirements, the reliability of the scoring system was investigated in an experiment, where six nurses scored 500 SIMP answers.

Interrater-reliability among these six nurses proved to be high. The correlation of one random selected nurse, with the mean of the population, was estimated .83 with Intra Class Correlation (ICC). Further analysis revealed some imperfections in the scoring-models, and suggested a lack of precision by one of the nurses. It was estimated that improvement of the scoring-models, and selection of raters on the characteristic of accuracy, could elevate the interrater reliability to .93. It was concluded that the scoring method may be regarded as objective.

Next the answers on four cases were also rated by two experienced physicians, with the same scoring-models. The overall agreement among all eight raters was again high. The inter-rater reliability was estimated .80. Only one case, with a defective scoring-model, showed a significant difference between the ratings by the nurses and by the physicians. Further analysis revealed, that the agreement among the nurses was stronger than that among the physicians. The relatively low agreement between the two physicians, may be explained by their expert knowledge. Their experience enables them to value the context of the whole answer, thereby the marking of items may be influenced. Differences in medical judgement than become visible in the scoring. Nurses, who do not have the expert knowledge to judge the context of an answer necessarily comply closer with the scoring-models. Despite some evidence that the nurses err in a few cases by lack of expert knowledge, it was concluded, that nurses meet the requirements of objective judgement better than physicians.

Reliability

Rating is only one of the facets that determine the reliability of a measure. Reliability is the degree to which a measure produces the same results, when it is repeated. In the repetition, facets like respondents, questions, and raters may be constant, or varied. Generalizability theory provides a framework that allows the analysis of all these facets at the same time.

The results of generalizability analyses on the data from the rater-reliability studies were disappointing. It was estimated that even a test of 30 cases would not reach satisfactory reliability.

However, in a next study a generalizability coefficient of .74 was found for a six-cases SIMP test, scored by four raters. This resulted in an acceptable reliability with a test consisting of 11 to 15 cases. The difference in result between the generalizability analyses can be explained partly by methodological problems. An alternative explanation is that the generalizability is domain specific.

Validity

The validity of SIMP was investigated by means of correlations with concurrent measures. Support for the validity of SIMP as a measure of medical problem-solving was found in a significant correlation with a global rating of performance with a simulated patient (SP). The insignificant correlation with assessment of medical knowledge by means of the Maastricht Progress Test was interpreted as circumstantial evidence of support. A high correlation would have been unfavorable, since that would mean that the problem-solving measure does not add information to the knowledge test.

The validity of SIMP was further supported by investigations of clinical competence, that applied SIMP as a measure. For instance, more detailed analyses of the relation between SIMP-scores and ratings with a simulated patient were reported by Crijnen et al (1987). Parts of SIMP were found to correlate highly with equivalent elements of their instrument for the assessment of medical interviewing skills.

In a study by Rethans and Van Boven (1987) scores on the written SIMP-test were compared with actual performance in medical practice. On the whole the results seemed equitable. Further analyses, however, revealed that the performance of physicians in practice was actually better than suggested by the written test.

The relation between SIMP-scores and scores on a knowledge test is further investigated by Van Leeuwen (1987). A distinct relationship could not be demonstrated. Van Leeuwen also reports a positive judgement on SIMP by the subjects in her investigation. The regard SIMP as a welcome extension to the objective knowledge tests.

With respect to further research on the validity of SIMP attention should focus on the question how to sample a representative test content (a test blueprint). Investigation of the relation of performance in actual practice could provide insight in the factors determining that performance. Also important is further exploration of the relation between SIMP-scores and factual medical knowledge. More insight in the structure of medical knowledge could provide clues toward the question whether

knowledge acts as a condition for problem-solving or if problem-solving ability is inherent in the structure of knowledge.

Conclusion

Application of SIMP in investigations on medical competence, as a concurrent or criterion measure seems justified.
On the whole, the research reported in this thesis supports the operationalization of the construct medical problem-solving with SIMP. It appears that reliable test can be constructed with 11 to 15 cases (a testing time of two and a half hours to three hours). There is, however, evidence suggesting that such a reliability can only be attained within a limited domain. Since that would imply that a much larger number of cases would be necessary for the assessment of "medical problem-solving", in general implementation of SIMP as an extension to the Progress Test seems inappropriate. As an alternative SIMP, or SIMP-like measures could be applied for assessment within the clinical clerkships. Analogous to the SIMP-test that was constructed for the family physician clerkship, versions could be adapted for almost any practical healthcare profession.
In such a way SIMP could contribute to the standardization of feed-back to students on their ability to handle practical problems in medical practice. Especially in a problem-based curriculum, aiming at preparation for practice, such feed-back is of utmost importance.

SAMENVATTING

Het meten van medisch probleemoplossen

Medisch probleemoplossen wordt vaak benadrukt als een belangrijke algemene doelstelling van medisch onderwijs. Een meting van vaardigheid in medisch probleemoplossen dient daarom deel uit te maken van de toetsing van medische studenten.
Aan de Rijksuniversiteit Limburg is voor het toetsen van medische kennis is een apart instrument ontwikkeld: de Maastrichtse Voortgangstoets. In aansluiting op deze Voortgangstoets bestond behoefte aan een instrument voor het meten van medisch probleemoplossen.
Bestaande instrumenten, zoals het Patient Management Problem (PMP) en de Modified Essay Question (MEQ) vertoonden echter tekortkomingen. Zo bleek een sturend effect van de antwoordopties de resultaten van het PMP te vertekenen. Verder gaf het regelmatig vinden van lage correlaties tussen casus onderling aanleiding tot twijfel ten aanzien van de validiteit van de metingen. Daarom werd besloten een nieuw instrument voor het meten van medisch probleem-oplossen te ontwikkelen.

Simulatie van Initieel Medisch Probleemoplossen (SIMP)

Onderzoek naar de aard van medisch probleemoplossen heeft uitgewezen, dat het vermogen direkt in het begin van het kontakt initiële hypothesen te formuleren van cruciaal belang is voor het verdere verloop van het kontakt. Daarom is bij de constructie van een nieuw instrument de nadruk gelegd op het begin van het arts-patiënt kontakt. Verder is gekozen voor open vragen, om het sturende effect van antwoordopties te vermijden.
Het op basis van deze uitgangspunten ontwikkelde instrument: Simulatie van Initieel Medisch Probleemoplossen (SIMP), bestaat uit korte beschrijvingen van casuïstiek, gevolgd door een enkele open vraag: "Wat zou u

doen, als u als arts in de praktijk met een dergelijk geval werd gekonfronteerd?" De respondenten zijn bij deze vraag vrij hun eigen formuleringen te kiezen. De antwoorden worden alleen beperkt tot de informatie die uit de gepresenteerde casus kan worden afgeleid. Feed-back wordt pas achteraf gegeven. De benodigde tijd per casus kan daarmee worden teruggebracht tot vijf à tien minuten. Binnen een toets kunnen daardoor tien tot twintig verschillende casus worden opgenomen.

Betrouwbaarheid van de beoordelingen

Een bezwaar van toetsing door middel van open vragen is de subjectieve invloed van de beoordelaars. Om dit effekt te ondervangen zijn voor de beoordeling antwoordsleutels ontwikkeld. Deze antwoordsleutels hebben de vorm van een checklist met omschrijvingen van elementen van een correct antwoord. Om het scoren te stroomlijnen zijn de items ingedeeld volgens het SOEP-schema van Weed (1969. De respondenten zijn echter vrij in het kiezen van een eigen struktuur voor hun antwoord. Van de beoordelaars wordt geen inhoudelijk waardeoordeel verwacht, maar alleen het markeren van overeenstemming tussen het antwoord en de antwoordsleutel. De beoordelaar hoeft dan ook geen medicus te zijn. Kennis van medische begrippen en terminologie is echter noodzakelijk om alternatieve formuleringen te kunnen herkennen. Aangezien verpleegkundigen op grond van hun opleiding en ervaring aan dit kriterium voldoen, is de betrouwbaarheid van het scoringssysteem onderzocht in een experiment, waarbij zes verpleegkundigen in totaal 500 antwoorden beoordeelden. De overeenstemming tussen deze beoordelaars bleek hoog te zijn. Met de Intraclass Correlatie Coefficient (ICC) werd de correlatie van beoordeling door één random gekozen verpleegkundige met het gemiddelde van de populatie verpleegkundige beoordelaars bepaald op .83. Nadere analyse bracht aan het licht, dat onvolkomenheden in enkele antwoordsleutels en slordigheid van een der beoordelaars de betrouwbaarheid negatief beïnvloedden. Geschat werd dat verbetering van de antwoordsleutels en selectie van beoordelaars de betrouwbaar-

heid verhoogd zou kunnen worden tot .93. Ten aanzien van de scoringsmethode werd gekonkludeerd, dat deze als objectief te beschouwen is. Vervolgens is onderzocht in hoeverre de beoordelingen van deze verpleegkundigen verschillen van beoordelingen door artsen, die beschouwd kunnen worden als inhoudelijke experts. Daarvoor zijn de antwoorden op vier casus opnieuw gescoord door twee artsen, met dezelfde antwoordsleutels. Over het geheel genomen was de overeenstemming tussen alle beoordelaars hoog. De beoordelaarsbetrouwbaarheid werd geschat op .80. Alleen bij een casus, waar al eerder manco's in de antwoordsleutel aan het licht gekomen waren, werd een significant verschil gevonden tussen de beoordelingen door de verpleegkundigen en die door de artsen. Nadere analyse bracht echter aan het licht dat de onderlinge overeenstemming tussen de twee artsen aanmerkelijk lager was dan die tussen de verpleegkundigen. De relatief lage overeenstemming tussen de twee artsen kan verklaard worden vanuit hun inhoudelijke deskundigheid. Doordat zij op grond van hun eigen ervaring een oordeel hebben over de kwaliteit van het gehele antwoord, kan dit oordeel mee van invloed zijn op het scoren van items. Verschillen in opvatting ten aanzien van medisch handelen kunnen dan tot uiting komen in de scoring. Verpleegkundigen, die niet een dergelijk eigen oordeel over de kwaliteit van het antwoord hebben, houden zich strikter aan de scoringsinstruktie. Ondanks aanwijzingen dat de verpleegkundigen, door gebrek aan inhoudelijke deskundigheid, in enkele gevallen tot een onjuist oordeel komen, kon daarom gekonkludeerd worden dat verpleegkundigen in dit aspekt beter voldoen als objektieve beoordelaars dan artsen.

Betrouwbaarheid

De beoordelaars vormen slechts een van de facetten, die van invloed zijn op de betrouwbaarheid van een meetinstrument. Betrouwbaarheid kan worden opgevat als de mate waarin een meting bij herhaling dezelfde resultaten oplevert. Facetten als respondenten, vragen en beoordelaars kunnen daarbij worden gevarieerd. Generaliseerbaarheidstheorie biedt een kader, waarin al deze facetten gelijktijdig geanalyseerd kunnen

worden. De resultaten van een generaliseerbaarheidsanalyse op basis van het materiaal van het beoordelaarsonderzoek waren teleurstellend. Een toets bestaande uit 30 casus was niet voldoende voor het realiseren van een acceptabele betrouwbaarheid. In een volgende studie werd echter voor een SIMP-toets bestaande uit zes casus, met vier beoordelaars een generaliseerbaarheidscoëfficiënt van .74 gevonden, resulterend in een acceptabele betrouwbaarheid bij 11 tot 15 casus. Voor een deel kan het verschil in uitkomst tussen de generaliseerbaarheidsstudies op methodologische gronden verklaard worden. Een alternatieve verklaring wordt gevormd door de mogelijkheid dat de generaliseerbaarheid domein specifiek is.

Validiteit

De validiteit van SIMP is onderzocht, door na te gaan hoe de toets correleert met concurrerende metingen. Steun voor deze concurrente validiteit van SIMP werd gevonden in de significante correlatie met een globale beoordeling van een Simulatie Patiënt kontakt (SP). De niet significante correlatie met de meting van medische kennis met behulp van de voortgangstoets, werd als indirekte ondersteuning geïnterpreteerd. Een hoge correlatie zou ongunstig zijn, aangezien dat zou betekenen dat de probleemoplostoets geen informatie aan de kennistoets toevoegt.

Verdere ondersteuning voor de validiteit van SIMP werd gevonden in onderzoeken van medische competentie, waarbij SIMP is toegepast als een van de meetinstrumenten. Zo werd de samenhang tussen SIMP-scores en beoordelingen van prestaties met een Simulatie Patiënt is nader geanalyseerd door Crijnen et al (1987). Onderdelen van de SIMP bleken hoog te correleren met overeenkomstige elementen van een instrument voor het beoordelen van medische gespreksvaardigheid.

De overeenstemming tussen feitelijke prestaties in de praktijk en meting met SIMP is onderzocht door Rethans en van Boven (1987). In grote lijnen stemden de metingen overeen. Bij nadere analyse bleken de prestaties van artsen in werkelijkheid beter te zijn dan gesuggereerd

door het antwoord op de schriftelijke simulatie. De ervaren artsen in dit onderzoek bleken niet alles op te schrijven wat ze wisten.
Verder is door Van Leeuwen (1987) onderzoek uitgevoerd naar de samenhang tussen resultaten van een meting met SIMP en een kennistoets. In dit onderzoek kon geen duidelijke relatie tussen kennis en probleemoplossend vermogen worden aangetoond. De lage betrouwbaarheid van de kennistoets kan echter een storende invloed hebben gehad. De betrouwbaarheid van de in dit onderzoek gebruikte SIMP toets was opnieuw bevredigend te noemen. Verder rapporteert Van Leeuwen over beoordeling van SIMP door de respondenten in haar onderzoek. Deze zijn over het algemeen zeer positief en zien SIMP als een welkome aanvulling op de gangbare objectieve kennistoetsing.
Wat betreft verder onderzoek naar de validiteit van SIMP verdient met name de vraag hoe een inhoudelijk representatieve toets moet worden samengesteld (een toets blauwdruk) nadere aandacht. Onderzoek naar de relatie met prestaties in de praktijk zou meer inzicht kunnen verschaffen, in de factoren die deze prestatie bepalen. Van belang is ook verdere exploratie van de relatie tussen SIMP en medische kennis. In de eerste plaats kan daarbij gedacht worden aan onderzoek naar de relatie tussen prestaties op SIMP en medische kennis. Daarnaast is echter ook onderzoek naar de struktuur van die medische kennis van belang. Hiermee kan worden nagegaan in hoeverre kennis fungeert als voorwaarde voor het kunnen oplossen van medische problemen, dan wel dat de probleemoplosvaardigheid verankerd is in de kennisstruktuur.

Conclusie

In het verlengde van eerdere onderzoeken waarbij SIMP gebruikt is als meting van medisch probleemoplossen, zijn er mogelijkheden voor toepassing van SIMP als concurrerende, of als criteriummeting.
Over het geheel genomen wordt de geldigheid van de operationalisatie van het construct medisch probleemoplossen met SIMP door het tot nu toe uitgevoerde onderzoek ondersteund. Gebleken is dat met deze methode betrouwbare toetsen kunnen worden samengesteld van 11 tot

15 casus (een testtijd van twee en een half a drie uur). Er zijn echter aanwijzingen dat een dergelijke betrouwbaarheid alleen gerealiseerd kan worden binnen een beperkt inhoudelijk domein. Aangezien voor meting van een algemene trek "medisch probleemoplossen" een veel groter aantal casus nodig zou zijn (met aanmerkelijk langere testtijd) ligt invoering van SIMP in het kader van de voortgangstoets niet voor de hand. Een goed alternatief is de toepassing van SIMP of SIMP-achtige meetinstrumenten bij de beoordelingen in de klinische stages. Naar analogie van de SIMP-toets voor het PMOH op het gebied van de huisartsengeneeskunde kunnen voor nagenoeg elke professie binnen de gezondheidszorg aangepaste versies worden ontwikkeld.
Op die wijze zou SIMP kunnen bijdragen aan standaardisering van de feed-back aan studenten ten aanzien van hun vermogen tot het oplossen van problemen in de medische praktijk. Het geven van dergelijke hoogwaardige feed-back is van groot belang, met name in een probleem-gestuurd curriculum waar voorbereiding op de praktijk sterk wordt benadrukt.

References

Basisfilosofie (1972) Rijksuniversiteit Limburg, Maastricht.

Barrows, H.S. (1971) Simulated Patients. Springfield Illinois, Charles C. Thomas.

Barrows, H.S. and Tamblyn, R.M. (1980) Problem-based Learnning. New York, Springer.

Barrows, H.S., Norman, G.R., Neufeld, V.R. and Feightner, J.W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. Clinical and Investigative Medicine, 5, 49-55.

Bender, W., Cohen-Schotanus, J., Imbos, T., Versfelt, W.A. and Verwijnen, M. (1984) Medische kennis bij studenten uit verschillende medische faculteiten: van hetzelfde laken een pak? Nederlands Tijdschrift voor de Geneeskunde, 128, 917-921.

Bligh, T.J. (1980) Written simulation scoring: a comparison of nine systems. American Educational Researcg Association, Annual Meeting, New York.

Bloom, B.S., Hastings, J.Th. and Madaus, G.F. (1971) Handbook on Formative and Summative Evaluation. New York, McGraw-Hill.

Boshuizen, H.P.A. (1989) De ontwikkeling van medische expertise: een cognitief psychologische benadering. Thesis, Maastricht.

Boshuizen, H.P.A. and Claessen H.F.A. (1982) Problems of research into medical problem solving: some remarks on theory and method. Medical Education, 16, 81-87.

Boshuizen, H.P.A., Schmidt, H.G. and Coughlin, L.D. (1988) On the application of medical science knowledge in clinical reasoning: Implications for structural knowledge differences between experts and novices. Paper presented at the 10th annual conference of the cognitive science society, Montreal, Canada.

Bouhuijs P.A.J., (1983). De ontwikkeling van het praktisch medisch onderwijs in de huisartspraktijk. Thesis, Maastricht.

Bouhuijs, P.A.J., Van der Vleuten, C.P.M. and Van Luyk, S.J. (1987) The OSCE as part of a systematic skills training approach. Medical Teacher, 9, 183-191.

Brennan, R.L. (1983) Elements of Generalizability Theory. Iowa City, Iowa, American College Testing Program.

Claessen, H.F.A. and Boshuizen, H.P.A. (1985) Recall of medical information by students and docters. Medical Education, 19, 61-67.

Conger, A.J. (1980) Integration and Generation of Kappa's for Multiple Raters. Psychological Bulletin, 88, 322-328.

Crijnen, A.A.M., Post, G.J., Kraan, H.F., Van der Vleuten, C., Imbos, T. and Zuidweg, J. (1987) Interviewing skills and medical competence. In: Kraan, H.F. and Crijnen, A.A.M., The Maastricht History-taking and Advice Checklist. Thesis, Amsterdam.

Cronbach, L.J., Glaser, G.C., Nanda, H. and Rajaratnam, N. (1972) The dependability of behavioral measurements. New York, Wiley.

Cutler, P. (1979) Problem-solving in clinical medicine: From data to diagnoses. Williams and Wilkins, Baltimore.

De Graaff, E. (1988) Simulation of Initial Medical Problem-solving: a test for the assessment of medical problem-solving. Medical Teacher, 10, 1, 49-55.

De Graaff, E. and Galesloot, J.A.M. (1981) Konstruktie en afname van een toetsmethode voor praktische medische kompetentie. Onderzoek van Onderwijs, 9, Rijksuniversiteit Limburg, Maastricht.

De Graaff, E. and Galesloot, J.A.M. (1982) De ontwikkeling van een toetsmethode voor 'medisch probleem-oplossen. In: H.G. Schmidt (red) Probleemgestuurd onderwijs. Harlingen, Flevodruk.

De Graaff, E., Moust, J.H.C., Ronteltap, C.F.M. and Schmidt, H.G. (1982) Studiebeleving van Maastrichtse medische studenten. In: H.G. Schmidt (red.) Probleemgestuurd onderwijs. Harlingen, Flevodruk.

De Graaff, E., Post, G.J. and Drop, M.J. (1987-a) Validation of a new measure of clinical problem-solving. Medical Education, 21, 213-218.

De Graaff, E., Drop, M.J., Post, G.J. and De Roos, K.P. (1987-b) Carrièrevoorkeuren en entree op de arbeidsmarkt van Maastrichtse basisartsen. Nederlands Tijdschrift voor de Geneeskunde, 131. 38, 1677-1678.

De Groot, A.D. (1969) Methodology. Foundations of inference and research in the behavioral science. The Hague, Mouton.

De Groot, A.D. (1973) Selektie voor en in het hoger onderwijs: Een probleemanalyse (Samenvatting deel I). Tijdschrift voor Opvoedkunde, 18, 364-373.

De Koning, P. (1977) Een oriënterende studie naar de afsluitingsproblematiek van de middenschool. Amsterdam, UvA, Pedagogisch Didactisch Instituut.

Drenth, P.J. (1971) De Psychologische Test. Deventer, Van Loghum Slaterus.

Ebel, R.L. and Frisbie, D.A. (1986) Essentials of Educational Measurement. (fourth edition) Englewood Cliffs, New Jersey, Prentice-Hall.

Elstein, A.S., Schulman, L.S. and Sprafka, S.A. (1978) Medical Problem-Solving: an analysis of clinical reasoning. Cambridge Massachusetts, Harvard University press.

Feightner, J.W. and Norman, G.R. (1976) Concurrent validity of patient management problems by comparison with the clinical encounter. Proceedings 15th Annual Conference on Research in Medical Education.

Feightner, J.W. and Norman G.R. (1978) Computerbased problems as a measure of the proclem-solving process. Some concerns about validity. Proceedings 17th Annual Conference on Research in Medical Education, New Orleans.

Feightner, J.W. (1985) Patient Management Problems In: Neufeld, V.R. and Norman G.R. (eds). Assessing Clinical Competence. New York, Springer Publishing Company.

Feletti, G.I. (1980) Reliability and validity studies on modified essay questions. Journal of Medical Education, 55, 933-41.

Feletti, G.I. and Smith E.M.K. (1986) Modified essay questions: are they worth the effort? Medical Education, 20, 126-132.

Fleiss, J.L. and Cuzick J. (1979) The reliability of dichotomous judgements: unequal numbers of judges per subject. Applied Psychological Measurement, 3, 537-542.

Frane, J. (1979) Analysis of the Pattern of Missing Data In: Dixon, W.J. and Brown, M.B. (eds.) BMDP-79. Berkeley, University of California Press.

Freankel, G.J. (1978) McMaster Revisited. Brittish Medical Journal, 2, 1072-1076.

Frederiksen, N. and Ward, W.C. (1978) Measures for the study of creativity in scientific problem-solving. Applied Psychological Measurement, 2, 1-24.

Frederiksen, N. (1984) The real test bias; Influences of testing on teaching and learning. American Psychologist, 3, 193-202.

Gale, J. (1981) Some cognitive components of the diagnostic thinking process. Brittish Journal of Educational Psychology, 52, 64-76.

Galesloot, J.A.M. and De Graaff, E. (1980) Definiëring van het begrip Medisch Probleemoplossen. Deelprojektgroep Summatieve Evaluatie, no. 13, Rijksuniversiteit Limburg, Maastricht.

Galesloot, J.A.M. and De Graaff, E. (1981) Casuïstiek uit de huisartsen praktijk; - 40 praktijksituaties voor zelfevaluatie. Utrecht/Antwerpen Scheltema en Holkema.

Galesloot, J.A.M., De Graaff, E., Imbos, Tj. and Verwijnen, M. (1981) Free-response test versus multiple choice tests. Medical Education, 15, 204-205.

Galofré, A. (1974) Review of written patient management simulations. Center for Educational Development, University of Illinois, Chicago.

Gerritsma, J.G.M. and Smal J.A. (1982) De werkwijze van huisarts en internist: een vergelijkend onderzoek met behulp van interaktieve patiëntensimulatie. Thesis, Utrecht.

Gerritsma, J.G.M. and Smal, J.A. (1988) An interactive patient simulation for the study of medical decision-making. Medical Education, 22, 118-123.

Goran, M.J., Williamson, W.J. and Gonella J.S. (1973) The validity of patient management problems. Journal of Medical Education, 48, 171-177.

Grant, I. and Marsden, P. (1987) The structure of memorized knowledge in students and clinicians: An explanation for diagnostic expertise. Medical Education, 21, 92-98.

Greep, J.M. (1979) Het onderwijs aan de medische faculteit in Maastricht. Medisch Contact, 35, 1107-1110.

Groen, G.J. and Patel, V.L. (1985). Medical problem-solving: some questionable assumptions. Medical Education, 95-110.

Guilford, J.P. (1954) Psychometric methods. New York, McGraw-Hill.

Harasym, P., Baumber, J., Bryant, H., Fundytus, D., Preshaw, R., Watanabe, M. and Wyse, G. (1980). An evaluation of the clinical problem-solving process using a simulation technique. Medical Education, 14, 381-86.

Hartmann, D.P. (1977) Considerations in the choice of interobserver reliability estimates. Journal of Applied Behavioral Analysis, 10, 103-116.

Hobus, P.P.M., Schmidt, H.G., Boshuizen, H.P.A. and Patel, V.L. (1987) Contextual information in the activation of first diagnostic hypotheses: expert-novice differences. Medical Education, 21, 471-476.

Hofstra, M.L., Hobus, P.P.M., Boshuizen, H.P.A. and Schmidt, H.G. (1988) Diagnostiek bij huisartsen: het belang van gegevens uit de context van de patiënt. Onderzoek van Onderwijs, 37, Rijksuniversiteit Limburg, Maastricht.

House, A.E., House, B.J. and Campbell, M.B. (1981) Measures of interobserver agreement: Calculation formulas and distribution effects. Journal of Applied Behavioral Analysis, 3, 37-57.

Kane, M.T. (1982) The Validity of Licensure Examinations. American Psychologist, 37, 911-918.

Kerlinger, F.N. (1973) Foundations of Behavioral Research. Tokyo, Holt-Saunders.

Knox, J..D.E. (1975) The Modified Essay Question. Univesity of Dundee.

Kratochwill, T.R. and Wetzel, R.J. (1977) Observer agreement, credibility, and judgement: Some considerations in presenting observer agreement data. Journal of applied behavior analysis, 10, 133-139.

Lamont, C.T. and Hennen, B.K.E. (1979) The use of simulated patients in a certification examination in family medicine. Journal of Medical Education, 47, 789.

Landis, R. and Koch, G.G. (1975) A review of statistical methods in the analysis of data arising from observer reliability studies (part 1). Statistica Neerlandica, 29, 101-123.

Lindquist, E.F. (1953) Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin Comp.

Linschoten, J.S. (1964) Idolen van de psycholoog. Utrecht, Beijleveld.

Maatsch, J.L., Munger, B.S. and Podgorny, G. (1982) Reliability and validity of the Board examination in emergency medicine. Emergency Medicine Annual, 1. Appleton, Century, Croft, Norwalk.

Martin, I.C. (1975) Emperical examination of the sequential management problem for measuring clinical competence. University of Illinois, Chicago.

Maxwell, A.E. and Pilliner A.E.G. (1986) Deriving coefficients of agreement for ratings. British Journal of Mathematical and Statistical Psychology, 21, 105-116.

McCorquodale, K. and Meehl, P.E. (1948) Operational validity of intervening constructs. In: M.H. Marx (ed.) Psychological Theory. New York, McMillan.

McGuire, C. (1973) Simulation technique in the teaching and testing of problem-solving skills. Columbus, Ohio, The Ohio State University.

McGuire, C. (1974) An overview of the use of simulation as an evaluation technique. In: Muslin, H.L. et al (eds.) Evaluative methods in psychiatric education. American Psychiatric Association, Washington.

McGuire, C. (1976) Simulation Technique in the teaching and testing of problem solving skills. Journal of research in science teaching, 13 (2), 89.

McGuire, C.H. and Babbot D. (1967) Simulation technique in the measurement of problem-solving skills. Journal of Medical Education, 4, 1-10.

McGuire, C.H., Solomon L.M. and Bashook P.G. (1972) Handbook of written simulations their construction and analysis. Center for Education Development, Chicago, Illinois.

Mellenbergh, G.J. (1977) The replicability of measures. Psychological Bulletin, 84, 378-384.

Mitchell, S.K. (1979) Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

Muller, S. (Chairman) (1984) Physicians for the twentyfirst century. Report of the Project Panel on the General Professional Education of the Physician and College Preparation for Medicine. Journal of Medical Education, 59, part 2.

Mura, E.L., Joorabchi, B. and Chawan, R. (1976) Medical competence: is the ECFMG examination a relevant measure? Journal of Medical Education, 51, 127-129.

Neufeld, V.R. and Barrows H.S. (1974) The McMaster philosophy: an approach to medical education. Journal of Medical Education, 49, 1040-1050.

Neufeld, V.R., Norman, G.R., Barrows, H.S. and Feightner, J.W. (1981) Clinical problem-solving of medical students: a longitudinal and cross-sectional analysis. Medical Education, 15, 26-32.

Neufeld, V.R. (1984) The design and use of assessment methods for problem-based learning. In: H.G. Schmidt, and M.L. De Volder (eds.) Tutorials in Problem-based learning: A new direction in teaching in the health professions. Assen/Maastricht, Van Gorcum.

Neufeld, V.R. and Norman G.R. (eds). (1985) Assessing Clinical Competence. New York, Springer Publishing Company.

Newble, D.I., Baxter, A. and Elmslie, R.G. (1979) A comparison of multiple-choice tests and free response tests in examinations of clinical competence. Medical Education, 13, 263.

Newble, D.I., Hoare, J. and Baxter, A. (1982) Patient Management Problems: issues of validity. Medical Education, 16, 137-142.

Newell, A. and Simon, H.A. (1972) Human Problem Solving. Englewood Cliffs (N.Y.) Prentice Hall.

Noll, V.H. (1965) Introduction to Educational Measurement. (sec. ed.) Houghton Mifflin, Boston.

Norman, G.R. (1985-a) Objective measurement of clinical performance. Medical Education, 19, 43-47.

Norman, G.R. (1985-b) Defining competence: a methodological review. in: Neufeld, V.R. and Norman G.R. (eds.) Assessing Clinical Competence. New York, Springer Publishing Company.

Norman, G.R. (1987) The status of research on evaluation of clinical competence. Paper presented at the International Symposium on Evaluation in Medical Education, Beer-Sheva, Israel, may 25-28.

Norman, G.R. (1988) Problem-solving skills, solving problems and problem-based learning. Medical Education, 22, 279-286.

Norman, G.R. and Feightner, J.W. (1981) A comparisom of behavior on simulated patients and patient management problems. Medical Education, 15, 26-32.

Norman, G.R., Neufeld, V.R., Walsh, A., Woodward, C.A. and McConvey, G.A. (1985) Measuring physicians' performances by using simulated patients. Journal of medical education, 60, 923-933.

Page, G.G. and Fielding, D.W. (1980) Performance on PMP's and performance in practice, are they related? Journal of Medical Education, 55, 529-37.

Palva, I.P. (1974) Measuring Clinical Problem Solving. British Journal of Medical Education, 52-56.

Post, G.J., Hellemons-Boode, B.P.S., Van der Heijden, P.F.A., De Graaff, E. and Drop, M.J. (1985) Medische kompetentie: een vergelijking tussen verschillende meetinstrumenten. Onderzoek van Onderwijs, 28, Rijksuniversiteit Limburg, Maastricht.

Post, G.J., De Graaff, E. and Drop M.J. (1986) Duur en numeriek rendement van de opleiding tot basisarts in Maastricht. Nederlands Tijdschrift voor de Geneeskunde, 130, 42, 1903-1905.

Post, G.J. and Drop, M.J. (1989) Perceptions of the content of the medical curriculum at the medical faculty in Maastricht. In: Innovation in Medical Education: An evaluation of its present status. T. Khattab, H.G. Schmidt, Z. Nooman and E. Ezzat (eds.) New York, Springer Publishing.

Post, G.J., De Graaff, E. and Drop M.J. (1988) Efficiency of a primary-care curriculum. Annals of Community Oriented Education, 1, 25-31.

Rabinowitz, H.K. (1987) The modified essay question: an evaluation of its use in a family medicine clerkship. Medical Education, 21, 114-118.

Reichenbach, H. (1938) Experience and prediction. Chicago, University of Chicago Press.

Rethans, J.J.E. and Van Boven, C.P.A. (1987) Simulated patients in general practice: a different look at the consultation. British Medical Journal, 294, 809-812.

Rimoldi, H.J.A. (1961) The test of diagnostic skills. Journal of Medical Education, 36, 73-79.

Robinson, S.A. (1977) Written simulated problems as measures of change in problem-solving skills. (dissertation), Ann Arbor, Michigan, University Microfilms International.

Schmidt, H.G. and Bouhuijs, P.A.J. (1980) Onderwijs in Taakgerichte Groepen. Utrecht, Het Spektrum.

Schmidt, H.G. (1983) Problem-based learning: rationale and description. Medical Education, 17, 11-16.

Schmidt, H.G., Dauphinee, W.D. and Patel, V.L. (1987) Comparing the effects of problem-based and conventional curricula in an international sample. Journal of Medical Education, 62, 305-315.

Schmidt, H.G., Hobus, P.P.M., Patel, V.L. and Boshuizen, H.P.A. (1987) Contextual factors in the activation of first hypotheses: Expert-novice differences. Paper presented at the AERA-conference, Washington, D.C.

Schouten, H.J.A. (1986) Nominal scale agreement among observers. Psychometrica, 51, 453-466.

Schumacher, C.F. (1973) Validation of the American board of internal medicine written examination. A study of the examination as a measure of achievement in graduate medical education. Annuals of Internal Medicine, 78, 131-135.

Schumacher, C.F., Burg, F.D. and Taylor, W.C. (1975) Computerization of a patient management problem examination to prevent retracing. British Journal of Medical Education, 9, 281-5.

Schwabbauer, M.L. (1975) Use of the latent image technique to develop and evaluate problem solving skills. Journal of Medical Technology, 41, 12, 457.

Shrout, P.E. and Fleiss, J.L. (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 2, 420-428.

Skakun, E.N. and McLaughlin R.S. (1978) The dimensionality of linear patient management problems. Proceedings 17th RIME-conference.

Spaulding, W.P. (1969) The Undergraduate Medical Curriculum Model: McMaster University. Canada Med. Ass. J. 100, 659-664.

Swanson, D.B., Norcini, J.J. and Grosso L.J. (1987) Assessment of clinical competence: written and computer-based simulations. Assessment and Evaluation in Higher Education, 12 ,3.

Thorndike, R.L. (1982) Applied psychometrics. Boston, Houghton Mifflin Comp.

Van Berkel, H.J.M. (1984) De diagnose van toetsvragen. Thesis, Centrum voor Onderzoek van het Wetenschappelijk Onderwijs, Amsterdam.

Van der Vleuten, C.P.M. and Van Luyk, S.J. (1985) Evaluating undergraduate training in medical skills. Paper presented at the symposium on the evaluation of innovative curricula for the health siences, Ismaila, Egypt.

Van Leeuwen, Y. (1987) Toetsing medisch probleem oplossen: een oplosbaar probleem? Rijksuniversiteit Limburg, Maastricht.

Van Leeuwen, Y., De Graaff, E. and Drop, M.J. (1987) The construction of Simulation of Initial Medical Problem-solving (SIMP). Workshop presented at the international symposium on evaluation in medical education, Beer Sheva, Israel, may 25-28.

Verwijnen, G.M., Imbos, T., Snellen, H., Stalenhoef, B., Pollemans. M., Van Luyk, S., Sprooten, M., Van Leeuwen, Y. and Van der Vleuten, C.P.M. (1982) The evaluation system at the medical school of Maastricht. Assessment and Evaluation in Higher Education, 7, 3, 225-244.

Verwijnen, G.M., Van der Vleuten, C.P.M. and Imbos, T.A. (1989) Comparison of an Innovative Medical School with Traditional Schools: An analysis in the cognitive domain. In: T. Khattab, H.G. Schmidt, Z. Nooman and E. Ezzat (eds.) Innovation in Medical Education: An evaluation of its present status. New York, Springer Publishing.

Vu, N.V., Barrows, H.S., Paiva, R.A.E. and Dawson-Saunders, B. (1984) The Medical Reasoning Aptitude Test. In: H.G. Schmidt and M.L. De Volder (eds.) Tutorials in Problem-based learning: A new direction in teaching in the health professions. Assen/Maastricht, Van Gorcum.

Wakefield, J. (1985) Direct Observation. In: Neufeld, V.R. and Norman G.R. (eds), Assessing Clinical Competence. New York, Springer Publishing Company.

Wakeford, R., Bashook, P., Jolly, B. and Tothman, A. (eds.) (1984) Directions in clinical assessment. Report of the first "Cambridge conference", Cambridge, England, 25 june-1 july 1984.

Walton, H.J. (1985) Primary health care in European medical education: a survey. Medical Education, 19, 167-188.

Weed, L.L. (1969) Medical records, medical education and patient care. Year book, Medical Publishers Inc. Chicago.

Wesdorp, H. (red.) with Blok, H., De Graaff, E., Wolowitsj-Schelvis, A. and Zijlmans, S. (1979) Studietoetsen en hun effecten op het onderwijs. SVO-reeks, 15, 's-Gravenhage, Staatsuitgeverij.

Wijnen, W.H.F.W. (1984) Student Assessment: Introduction. In: H.G. Schmidt and M.L. De Volder (eds.) Tutorials in Problem-based learning: A new direction in teaching in the health professions. Assen/Maastricht, Van Gorcum.

Wijnen, W.H.F.W. and Van der Vleuten C.P.M. (1985) Toetsing: Hordenloop of Voortgangscontrole? Universiteit en Hogeschool, 31, 6.

Winer, B.J. (1971) Statistical principles in experimental design, 2nd ed. New York, McGraw-Hill.

Curriculum Vitae

Erik de Graaff was born on april 21th 1951 in Amsterdam. He graduated (HBS-b) at the Amsterdam Lyceum in 1970. He studied psychology at the University of Amsterdam and graduated as organizational psychologist (arbeids- en organisatie psychologie) in 1978.
From 1977 to 1979 he was employed as educational researcher at the Research Instituut voor de Toegepaste Psychologie in Amsterdam.
Since 1979 he is a member of the department Educational Research and Development (Onderwijsontwikkeling en Onderwijsresearch) of the Rijksuniversiteit Limburg. First as project researcher at the medical faculty, since 1981 as University Teacher (Universitair Docent) at the faculty of Health Sciences.